


For Release 2001/11/07 : CIA-RDP00-01458R000100110001-9



**PRINCIPLES FOR THE
VALIDATION AND USE OF
PERSONNEL SELECTION
PROCEDURES:
SECOND EDITION**

For Release 2001/11/07 : CIA-RDP00-01458R000100110001-9

This document is an official policy statement of the Division of Industrial-Organizational Psychology, American Psychological Association. It does not, however, necessarily represent the policy of the Association. Copies are available from the Secretary-Treasurer of the Division. The price schedule is:

\$4.00 each for 1-9 copies
\$2.50 each for 10-49 copies
\$2.00 each for 50 copies and up

Orders should be sent to:

Dr. Lewis E. Albright
Kaiser Aluminum & Chemical Corporation
300 Lakeside Drive—Room KB 2140
Oakland, California 94643

Published by *The Industrial-Organizational Psychologist*, Berkeley, California, and Printed by the University of California Printing Department, Berkeley, California.

Copyright
Division of Industrial-Organizational Psychology
1980

Citation: American Psychological Association, Division of Industrial-Organizational Psychology. *Principles for the validation and use of personnel selection procedures*. (Second edition) Berkeley, CA: Author, 1980.

Approved For Release 2001/11/07 : CIA-RDP00-01458R000100110001-9

**Principles for the Validation and
Use of Personnel Selection Procedures:
Second Edition**

**Division
of
Industrial-Organizational Psychology
American Psychological Association
1980**

Approved For Release 2001/11/07 : CIA-RDP00-01458R000100110001-9

Foreword

At the August 1978 meeting of the Division 14 Executive Committee the president, C. Paul Sparks, was instructed to appoint editors and an advisory panel to revise and update the *Principles for the Validation and Use of Personnel Selection Procedures* (1975). The reasoning behind this instruction included both the increased attention to tests and testing during the 1975-1978 period and a forecast for even greater attention in the future. This document resulted.

William A. Owens, Jr. and Mary L. Tenopyr accepted an invitation to serve as co-editors for the revision. Twenty-six Division members were invited to serve on the advisory panel. Twenty-five accepted (one later withdrew for personal reasons). The revision process was begun with a request that the advisory panel members furnish the co-editors with critical comments on the 1975 *Principles*. On the basis of these comments, a first draft was prepared and circulated to the advisory panel members. The responses were many and varied. Analysis of these indicated that major rewriting was necessary, not merely an update of the 1975 *Principles*. The target date had to be extended and the Executive Committee of Division 14 at its September 1979 meeting instructed the incoming president, Mary L. Tenopyr, to press forward. The Executive Committee also expressed a desire that every member of Division 14 have an opportunity to express her/his opinion before publication.

In December 1979, a draft was mailed to every Division 14 member, using mailing labels purchased from APA. In addition, addresses of new Division members not yet on APA rolls were secured, and they also received copies. In addition to a copy of the draft, each member received a questionnaire which asked for a rating of each section of the draft for agreement and clarity. A discussion of the analysis afforded the replies, and the results of the questionnaire were published in the May 1980 issue of *The Industrial-Organizational Psychologist*.

In April 1980 what was perceived as a final draft was mailed to all members of the advisory panel and to all members of the Executive Committee. With minor editorial revisions, this draft was presented to the Executive Committee at its May meeting. Publication was approved unanimously. This document is, therefore, an official document of the Division of Industrial-Organizational Psychology.

The Division is deeply indebted to the co-editors, the members of the advisory panel, and the membership at large for their constructive suggestions.

C. Paul Sparks, *President 1978-79*
Mary L. Tenopyr, *President 1979-80*

Executive Committee, Division 14

Lewis E. Albright, Ph.D.
Kaiser Aluminum & Chemical Corporation

Milton R. Blood, Ph.D.
Georgia Institute of Technology

Richard J. Campbell, Ph.D.
American Telephone and Telegraph Company

Milton D. Hakel, Ph.D.
The Ohio State University

Virginia E. Schein, Ph.D.
The Wharton School, University of Pennsylvania

Frank L. Schmidt, Ph.D.
Personnel Research & Development Center
U.S. Office of Personnel Management

Benjamin Schneider, Ph.D.
Michigan State University

C. Paul Sparks
Exxon Company, U.S.A.

Mary L. Tenopyr, Ph.D.
American Telephone and Telegraph Company

Paul W. Thayer, Ph.D.
North Carolina State University

Victor H. Vroom, Ph.D.
Yale University

Kenneth N. Wexley, Ph.D.
University of Akron

**Advisory Panel on Validation and Use of
Personnel Selection Procedures, Division 14**

William A. Owens, Jr. Ph.D. (Co-chair) University of Georgia	Edwin A. Fleishman, Ph.D. Advanced Research Resources Organization
Mary L. Tenopyr, Ph.D. (Co-chair) American Telephone and Telegraph Company	Donald L. Grant, Ph.D. University of Georgia
Lewis E. Albright, Ph.D. Kaiser Aluminum & Chemical Corporation	Robert M. Guion, Ph.D. Bowling Green State University
Philip Ash, Ph.D. University of Illinois, Chicago Circle	James J. Kirkpatrick, Ph.D. California State University, Long Beach
Richard S. Barrett, Ph.D. Consultant	Hobart Osburn, Ph.D. University of Houston
C. J. Bartlett, Ph.D. University of Maryland	Charles A. Pounian, Ph.D. City of Chicago
Brent N. Baxter, Ph.D. American Institutes for Research	Erich P. Prien, Ph.D. Memphis State University
Virginia R. Boehm, Ph.D. Sohio	Frank L. Schmidt, Ph.D. Personnel Research & Development Center U.S. Office of Personnel Management
William C. Burns Pacific Gas & Electric Company	Paul W. Thayer, Ph.D. North Carolina State University
Joel T. Campbell, Ph.D. Educational Testing Service	George C. Thornton, III, Ph.D. Colorado State University
Jerome E. Doppelt, Ph.D. The Psychological Association	Harold J. Tragash, Ph.D. Xerox Corporation
Marvin D. Dunnette, Ph.D. University of Minnesota	Kenneth N. Wexley, Ph.D. University of Akron
Frank W. Erwin Richardson, Bellows, Henry & Co., Inc.	Sheldon Zedeck, Ph.D. University of California, Berkeley

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 Division of Industrial-Organizational Psychology (Division 14) of the American Psychological Association as the official statement of the Division concerning procedures for validation research and personnel selection. Its purpose is to specify principles of good practice in the choice, development, and evaluation 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, lie detector or stress analyzer techniques, assessment center evaluations, biographical data forms or scored application blanks, scored or rated interviews, educational requirements, experience requirements, reference checks, physical requirements such as height or weight or physical ability testing devices, appraisals of job performance, estimates of advancement potential, or any other selection standard, whenever any one or a combination of these is used or assists in making a personnel decision.

When any selection procedure is used, the essential principle is that evidence be accumulated to show a relationship between decisions based on assessments made by that procedure and criteria such as job performance, training performance, advancement, or other pertinent job behavior.

This document is a revision of the *Principles* published in 1975 by that year's Division 14 Executive Committee. The revision was stimulated by ever-increasing attention to selection practices of employers. This attention has been made manifest by significant researches and theoretical formulations of measurement psychologists, by more detailed guidelines from equal employment opportunity enforcement agencies, and by numerous and diverse interpretations of the federal courts with respect to the extent to which the operational use of selection procedures comports with regulatory requirements and/or professional standards.

This statement intends to provide:

- (1) principles upon which the conduct of personnel research may be based,
- (2) guidance for practitioners conducting validation studies,
- (3) principles for application and use of valid selection procedures, and
- (4) information which may be helpful to personnel managers and others responsible for authorizing or implementing validation efforts.

The interests of some people will not be addressed by this statement. These *Principles* are not intended to:

- (1) be a technical translation of existing or anticipated regulation,
- (2) substitute for adequate training in validation procedures,
- (3) be exhaustive (although they cover the major aspects of validation), or
- (4) 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 predictive strategy. All measures made by a selection procedure are secured with express or implied expectation that they will be related to one or more important aspects of job behavior.

The principles presented here are generally stated in the context of traditional approaches. Other developments in validation research are addressed as appropri-

are but are not systematically developed here; e.g., the use of what has been described as *formal decision theory* (Cronbach & Gleser, 1965; Dunnette, 1974), the various forms of *synthetic validity* (Guion, 1965; McCormick & Mecham, 1970; Primoff, 1972), *Bayesian inference* (Novick & Jackson, 1974; Schmidt, Hunter, Pearlman, & Shane, 1979), or *internal/external validity* (Cook & Campbell, 1976, 1979; Cronbach, 1980). The traditional approaches are used as a framework because their concepts have been established through a long history and are explicated in most current text books. It is sometimes difficult to define "a long history." Two well-known and respected professionals may disagree vehemently as to whether a given position has been thoroughly established or is still in the developmental stage.

The *Principles* are not meant to be at variance with the *Standards for Educational and Psychological Tests* (APA, 1974). However, the *Standards* were written for measurement problems in general while the *Principles* are addressed to the specific problems of decision making in the areas of employee selection, placement, promotion, etc. In addition, a Joint Committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education has completed a review of the 1974 *Standards* and has recommended that they be revised, generally for the same reasons that this *Principles* revision was undertaken (AERA, APA, & NCME, 1979). Further, the Committee recommends, "The new *Standards* should be a statement of technical requirements for sound professional practice and not a social action prescription." This *Principles* revision is consistent with that expression.

Like the *Standards*, the *Principles* stated here present ideals toward which the members of this Division and other researchers and practitioners are expected to strive. Circumstances in any individual study or application will affect the importance of any given principle. Researchers and practitioners should, however, consider very carefully any factors suggesting that a general principle is inapplicable or that its implementation is not feasible. It is most appropriate to bear in mind the following statement from the *Standards*, cited in full in the 1975 *Principles* and now repeated here:

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 (APA, 1974, p. 8).

The *Principles* are intended to represent the consensus of professional knowledge and thought as it exists today, albeit not a *consensus omnium* since this is probably unattainable. Also, it is to be noted that personnel selection research and development is still an evolving field and techniques and decision-making models are subject to change. This document contains references for further reading and for support of the principles enunciated. It is expected that both researchers and practitioners will maintain an appropriate level of awareness of developments in the field.

Definition of Validity

Validity is the degree to which inferences from scores on tests or assessments are justified or supported by evidence. It should be noted that validity refers to the inferences made from the use of a procedure, not to the procedure itself. The

primary question to be answered in validation is the degree to which these inferences are appropriate. Use of a specific procedure may lead to valid inferences in one area and yet fail to lead to valid inferences in another area. It is incumbent on the investigator to define, in advance of any validation effort, the inferences to be made and to plan the validation strategy accordingly.

In planning validation it is not appropriate to think of validity as a single number or other result of a set of procedures. Several authors (e.g., Dunnette & Borman, 1979) have criticized the rigidity with which validation procedures have been applied, with apparently little thought of the meaning to be imparted to the results of the tests or other assessment procedures. A particular problem is the compartmentalization of validity into the categories of criterion-related, content, and construct. The three are really inseparable aspects of validity, not discrete types of validity. Although the three may represent differences in strategy, they do not necessarily indicate differences in concept. For example, aptitude tests are typically associated with criterion-related validation. In their development, items or components are frequently chosen on the basis of content sampling. Construct considerations are usually a major factor in defending the domain from which the items or components are sampled. Also, as mentioned earlier, prediction is often thought of as closely associated with criterion-related validation. In employment situations the use of scores from a procedure developed on the basis of content also has a predictive basis. That is, one measures performance in a domain of job activities which will be performed later. Furthermore, constructs may be said to underlie all predictions and so render score interpretations meaningful.

The *Principles* discuss these three validity strategies separately only to take advantage of traditional presentations. However, the reader is advised that in concept, and many times in methodology, the three cannot be logically separated. The *Principles* also use the term "strategy" instead of "validity" in labeling the three aspects. The purpose of this usage is to emphasize again the interrelatedness of the three aspects. 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 for practical considerations, but also to further the search for establishment of meaning relative to selection measures.

A Comment on "Fairness"

Social and legal influences have led to a concern, shared by psychologists, for *fairness* or *equality* in employment opportunity. A basic assumption of the principles of good practice is that those who follow them will also further the principle of fair employment. The interests of employers, applicants, and the public at large are best served when selection is made by the most valid means available. These *Principles* are technical in focus. They are primarily concerned with validity. The maximization of opportunities for *each individual* can be most effective where validity enables one to attain the highest level of accuracy in prediction or assessment of qualifications.

Fairness of a selection procedure, when criterion-related methodology is used, has been subject to many definitions. There are two basic classes of definition, the psychometric and the decision-making. The psychometric models advanced are numerous (Cleary, 1968; Cole, 1973; Darlington, 1971; Einhorn & Bass, 1971; Guion, 1966; Linn, 1973; Thorndike, 1971). Results yielded by these models are often not consistent with each other and may even be contradictory. The *Principles* do not at this stage of the professional debate advocate any one model. However, the reader is directed to Petersen and Novick (1976) who have pointed out that all of the models except those of Cleary and Einhorn and Bass have problems relative to their internal consistency. The model proposed by Guion is also not internally faulty. Most of the recent work has been devoted to decision-theoretic models (Cronbach, 1976; Petersen & Novick, 1976; Schwartz, 1978). These models require

an advance specification of utilities, thereby essentially removing the question of fairness from the hands of the psychometricians.

Application of Principles

It is not likely that anyone will completely satisfy the *ideal* of every applicable principle. This probability raises the question of relative levels of stringency in adhering to the individual principles. 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. In the operational use of validated selection procedures, the importance of adherence to the *Principles* again depends on the consequences of error. Will selection errors result in physical, psychological, or economic injury to people? Will the safety or operating efficiency of the organization be impaired because of selection errors? If so, then the principles may need to be followed more rigorously than in less crucial situations. Three axioms underlie the application of all these principles:

- (1) Individuals differ in many ways.
- (2) Individual differences in personal characteristics and backgrounds are often related to individual differences in behavior on the job.
- (3) It is in the best interest of organizations and employees that information about relevant differences between individuals be developed and used in assigning people to jobs.

Objectives of Validation Efforts

Before any assessment procedure is considered, or any validation effort is planned, one should have a clear idea of the objective of the assessment or validation. Any such statement of purpose logically must come from an understanding of the needs of the organization and of its present and prospective employees. As a general matter, a researcher should develop clear objectives for the proposed assessment procedure(s) and design the validation effort to determine how well they have been achieved. Objectives should be consistent with professional, ethical, and legal responsibilities.

Ideally, all aspects of the decision-making process should make a valid contribution to achievement of those objectives. Researchers should present evidence for the validity of as many aspects of the decision-making process as feasible. All assessment methods used should make a contribution to validity in ways which can be demonstrated. However, when it is impossible or infeasible to apply validation methods to a given part of the decision-making process, that part should have a relationship, discernible by a knowledgeable person, to appropriate purposes of the organization.

Job Analysis

A systematic examination of the job and the context in which it is performed will provide an enhanced understanding of the selection problem. This will also enhance the likelihood of finding a significant relationship between predictors and criteria in a criterion-related study through development of hypotheses concerning predictors and development or evaluation of criteria. Job analysis is essential to the development of a content oriented procedure. A number of job analysis procedures exist, each differing in terms of its possible contribution to the objectives of the particular study or a portion of the study (McCormick, 1979). There is currently no authoritative set of principles for job analysis comparable to the *Standards* or *Principles* in the area of selection procedures. The development of such a set is beyond the scope of this document. Discussed below are some of the elements of current practice and some of the constraints which they impose.

All formal job analysis techniques specify the *descriptors*, or units of analysis.

by which the job(s) will be defined. One way of classifying such techniques is by the nature of the descriptor specified and the type of job definition produced. For example, task analysis specifies the use of *task* or *activity statement* descriptors which culminate in a definition of the *job-oriented content* of the job(s); work behavior analysis specifies the use of *behavior statement* descriptors which culminate in a definition of the *worker-oriented content* of the job(s). Another way of classifying job analysis techniques is through the systems and methods used. Some systems provide a standardized set of job descriptors, usually an inventory or a questionnaire, which is programmed to provide output along a prescribed set of dimensions (Baehr, 1971; McCormick, Jeanneret, & Mecham, 1972; Pass & Cunningham, 1977). Other systems or methods require origination and development of the job descriptors by the analyst but with the analysis programmed to provide results according to a prescribed matrix of dimensions (Christal & Weissmuller, 1976; Fine & Wiley, 1971; Primoff, 1971). A summary of job analysis results to that date has been published by Prien and Ronan (1971).

The objective of the research is to obtain job information appropriate to the purpose or application of that job analysis information. The choice of job analysis methodology (e.g., the descriptors chosen and the job analysis operations used) is determined by that objective but with situational constraints. Constraints which need to be considered in the choice of method include, among others, the nature of the jobs, the situation, the resources available to the researcher, the research design and the types of evaluative operations which are included in the research design. For example, the extent to which the researcher's 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) reviews the literature and examines the conceptual and research issues in this area.

Criterion-Related Strategy

In general, the use of any personnel selection procedure is to predict future performance as measured by some job relevant *criterion*. Evidence for criterion-related validity typically consists of a demonstration of a statistically significant relationship between the selection procedure (predictor or predictors) and one or more measures of job relevant performance (criterion or criteria). It is, therefore, vital that the choice of both predictors and criteria be evaluated with great care.

In this section the word "*predictor*" will be used to refer to any aid to decision-making used in the context of personnel selection (in or out), placement, classification, or promotion. It will include, but not be limited to, standardized ability tests, personality inventories, biographical data forms, situational tests, assessment center evaluations, interview-based ratings, performance ratings, evaluations of training or experience, etc. (See Statement of Purpose, p. 1.) Predictors which are objective or "*standardized*" are preferred; i.e., where standard directions and procedures for administration, scoring, and interpretation are both delineated and employed. The principles of this section apply to all predictors, but more easily to those more rigorously standardized.

A. *Determination of Feasibility.* Anyone 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 a reasonably competent study; and although it may be argued that most errors merely reduce estimated validity, a poor study is *not* better than none. Several considerations are relevant in determining feasibility.

First, one must be able to assume that the job is reasonably stable and not in some period of rapid evolution. Although validity coefficients seem to be quite robust across both tasks and situations (Schmidt, Hunter, & Pearlman, in press), the traditional logic of validation research is that it is undertaken under conditions as comparable as possible to those which will exist when the results are made

operational. If this assumption is obviously and grossly in violation, it is incumbent on the researcher either to modify the validation strategy appropriately or to postpone the study until reasonable stability has returned.

Second, it must be possible to obtain or develop a *relevant*, reasonably *reliable* and *uncontaminated* criterion measure(s). Of these characteristics, the most important is relevance. This means that the criterion must accurately reflect the relative standing of employees with respect to prescribed job behaviors. If such a criterion measure does not exist or cannot be developed, criterion-related validation is not feasible. Criterion-related studies based upon criterion availability alone, rather than upon relevance, are inappropriate.

Third, a competent criterion-related validation should be based to the extent feasible on a sample which is reasonably representative of the populations of people and jobs to which the results are to be generalized. As mentioned previously, 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, which may obviously distort an estimate obtained from a particular sample. When there is evidence that gross distortion has occurred, the researcher must either estimate its impact, and adjust for it, or must conclude that it is not feasible to conduct a criterion-related validation.

Fourth, to conduct a criterion-related validity study which potentially lacks adequate "*statistical power*" may leave the issue of validity unresolved. The term *power* refers to the probability of obtaining a statistically significant relationship between predictor and criterion in a sample *if* such a relationship does, in fact, exist. Factors determining statistical power include sample size, degree of predictor range restriction, criterion reliability, and the size of the predictor-criterion relationship (Cohen, 1977). Combinations of these variables leading to low power can occur frequently in practice (Schmidt, Hunter, & Urry, 1976). As a consequence, it is quite possible to conclude that a significant predictor-criterion relationship is lacking when one does, in fact, exist. If the requirements cannot be met, the situation may not lend itself to a criterion-related validation.

Fifth, the previous discussion has implied the use of correlational statistics developed from predictor-criterion relationships. A special case must be made for those situations in which some intervening variable has essentially eliminated all variance from the criterion. An example of this is a self-paced training situation in which all selectees have attained a mastery level of the knowledge or skill being taught. If the training content is truly job related, no significant correlation can be obtained between training success and job success since there will be little or no variance in either the training success predictor or the job success criterion. There will be no significant correlation with a predictor selected to predict success in training since there will be little or no variance in the success in training criterion. Training time may be the only feasible criterion with an acceptable amount of variance present. Use of experimental and control groups with the experimental group selected on the predictor and the control group selected by some method which assures randomness may provide evidence of validity in such a situation (Goldstein, 1980).

B. Design and Conduct of 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 such a study. There are two criterion-related designs for generating evidence as to the validity of a measuring device.

One design employs the *predictive model* in which predictor information is obtained prior to placement of employees on a job and criterion information is obtained later. This design answers the most common employment question; i.e., does the predictor indeed have forecasting value with respect to *later* job behavior? As such, the predictive model addresses itself to the basic selection issue as it normally occurs in the employment context.

The other design is the *concurrent* model in which both predictor and criterion information are obtained for present employees at approximately the same time. The research literature clearly indicates that well conducted concurrent studies can provide useful estimates of predictive validity (Bemis, 1968; Pearlman, Schmidt, & Hunter, in press). Both types of criterion-related studies are susceptible to the effects of range restriction. However, the test scores obtained in concurrent studies may also be influenced by additional job knowledge, different motivation, or added maturity of incumbents vs. applicants. A concurrent study with appropriate controls should yield results very comparable to those of a predictive study.

1. *Criterion Development.* Once a validation model has been selected, the researcher should next be concerned with obtaining any necessary job information. In general, if criteria are chosen to represent job relevant activities or behaviors, the results of a formal job analysis will be helpful in criterion construction. Although numerous procedures are available (see p. 4), there does not appear to be a clear choice of method. What *is* essential, however, is that information about the job be competently and systematically developed. If the goal of a given study is the exclusive prediction of such nonperformance criteria as tenure or absenteeism, a formal job analysis will not usually be necessary, though an understanding of the job and its context will still be 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 on the basis of 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, it is not acceptable practice to substitute measures which are unrelated to the purposes of the study. One may not achieve the appearance of broad coverage by substituting irrelevant criteria which are available for relevant criteria which are unavailable.

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 determining what is and what is not included in a criterion. 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 researcher, 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).

c. *The Possibility of Bias or Other Contamination Should be Considered.* Although a simple group difference on the criterion does not *establish* bias, such bias would result if a definable subgroup were rated consistently and *spuriously* high (or low) as compared to other groups. Conversely, if a group difference did, in fact, exist but were not revealed by appropriate ratings, this would also constitute bias. It is therefore apparent that the presence or absence of bias cannot be detected from a knowledge of criterion scores alone. If objective and subjective criteria disagree, bias in the more subjective measure may be suspected, although bias is not limited to subjective measures. There is no clear path to truth in these matters. A criterion difference between older and younger employees, or day and night shifts may reflect bias in raters, equipment, or conditions, or it may also reflect genuine differences in performance. What is required is the anticipation and reduction of the possibility of bias, alertness to this possibility, protection against it insofar as is feasible, and use of the best judgment possible in evaluating

the data. Contamination, per se, could exist if selection test results were available to supervisors making presumably independent performance ratings. Correction after the fact is a near impossibility in this case.

d. *If Evidence Recommends that Several Criteria be Combined to Obtain a Single Variate, There Should be a Rationale to Support the Rules of Combination.* For example, it is probably generally preferable to weight for relevance, although special circumstances may occasionally argue otherwise. Thus, if well informed judges are unavailable, it may be best to assign unit or equal weights to the several criterion components.

e. *It is Desirable, But Not Essential, That Criterion Measures be Highly Reliable.* Reliability should be estimated, where feasible, and by appropriate methods (e.g., Stanley, 1971). It must 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.* There are numerous factors other than availability which 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. A study is more likely to indicate validity if there is a good reason to suppose that a relationship exists between a predictor chosen and the behavior it is supposed to predict. For example, the research literature or the logic of development may provide the reason. This principle does not intend to rule out application of serendipitous findings, although such findings usually need verification.

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 yield to the findings of relevant research and resist the influence of personal interest, mere familiarity or expediency. On the other hand, the researcher must exercise some 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. *Other Things Equal, Predictors Which Are More Objective Are to be Preferred.* Thus, the assessment of a candidate should be maximally dependent on his/her personal characteristics and minimally dependent on *who* made the assessment. Similarly, where non-test predictors like interviewer judgments are utilized, an effort should be made to develop procedures which will minimize such sources of error variance as are represented by differences between judges.

d. *Outcomes of Decision Strategies Should be Recognized as Predictors.* It must be noted that the decision-maker who interprets and acts upon a complex of predictor data interjects something of himself/herself into the interpretive or decision-making process. These judgments or these decisions thus become at the least an additional predictor, or at the most the *only* predictor. So, 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.

3. *Choice of Sample.* The meaningfulness of the research result is greatly dependent on the sample. Having several hundred subjects may not be better than one hundred if the selection of subjects chosen to obtain the larger N does not have an appropriate rationale.

a. *The Sample for a Validation Study Should be Carefully Chosen.* Whether the study is predictive or concurrent, the incumbent sample 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 character-

istics studied to date appear to have little or no effect on predictor-criterion relationships, no variable should be assumed to moderate validities in the absence of explicit evidence for such an effect. For example, the research literature shows that validities within 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 which 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 appropriate persons.

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.

d. *When Combining Data from Separate Samples, Both Jobs and Workers Should be Comparable on Variables Which Research has Shown to Affect Validity.* If comparability exists on these variables, pooled samples may be expected to provide increased statistical power.

4. *Procedural Considerations.* The researcher must consider the probable use of any end products. This should be done in advance of the collection and analysis of data.

a. *Validation Research Should Ordinarily 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 are *not* advanced to the higher level job, it may still be acceptable to evaluate 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. Predictability may diminish over long time spans as a result of changes in abilities and skills required, changes in the job itself, increased restriction of range in the subject pool, and related factors. On the other hand, predictability may increase as the demands of the higher level job result in greater differentiation of the performance of job incumbents or rate of advancement results in varying demands on the individual. Here again, the purposes of the study are paramount.

b. *Where Traditional Criterion-Related Validation Strategy is Not Feasible, the Researcher Should Consider Any Alternative Research Methodology Which Offers a Sound Rationale.* Examples include synthetic validation, cooperative research on an industry-wide basis, consortia of small users, or gathering data for validity generalization. However, the researcher should be aware that most non-traditional approaches require considerable research and development effort.

c. *Procedures for Test Administration and Scoring in Validation Research Should be Clearly Set Forth and Should be Consistent with the Standardization Planned 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. *It is Desirable That There 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 employer's need to get on with the business of making employment decisions. Where there is external evidence which supports the probability of valid prediction, it may be feasible to utilize the predictors immediately. However, the researcher must avoid situations that make it impossible or difficult to detect validity. For example, decisions should not be so highly selective that severe restriction of range results. If there is no firm basis for the presumption of validity, the researcher must carefully judge whether the dangers of postponing the use of the predictor are greater or less than the dangers of using it prematurely.

e. *The Collection of Predictor Data and Criterion Measures Should be Operationally Independent.* A common example of non-independence is the collection of criterion ratings from supervisors who know selection test scores. If a significant validity coefficient is obtained, it may be due either to a true relationship or to the manipulation of ratings (consciously or unconsciously) to conform with scores. Such ambiguity should be avoided.

5. *Data Analysis.* Modern computer technology allows the researcher to investigate different predictor-criterion relationships, different statistical techniques, etc., with considerable freedom and little cost. Any result based upon an extensive post hoc analysis should be replicated.

a. *The Method of Analysis Should be Chosen with Due Consideration for the Characteristics of the Data and the Assumptions Involved in the Development of the Data Analysis Method.* Some violations of assumptions can be tolerated with few ill effects; violations of others may produce grossly misleading results. It is the responsibility of the investigator to know the assumptions of the methods chosen and the consequences of violations of them.

b. *The Type of Statistical Analysis to be Used Should be Considered in Planning the Research.* The kinds of decisions to be made, and the way in which predictor variables are to be used in determining these decisions, should be considered in selecting the method(s) of analysis to be employed. Although any standard method(s) may be used, any new or unusual method should be clearly explained in the research report. (It is understood that conditions may develop in the course of an investigation which will require a change in plans.)

c. *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 a relationship. Traditionally, a validity coefficient or similar statistic which has a probability of less than one in twenty of having occurred by chance may be considered as establishing significant validity. There may be exceptions to this rule; professional standards have never insisted on a specific level of significance. However, departures from this convention should be based on reasons which can be stated in advance (such as power functions, utility, economic necessity, etc.). The analysis should provide information about the strength of the relationship. This is usually expressed in terms of coefficients of correlation but other methods (such as the slope of the regression line or the percentage of misclassifications) are acceptable and even preferable in many situations. The analysis should also give information about the nature of the relationship and how it might be used in prediction. For example, in comparing groups, the slope of the regression line is generally preferable to the coefficient of correlation. Use of expectancy tables may also be appropriate. Information provided should in any event include numbers of cases and measures of central tendency and variability for both predictor and criterion variables.

d. *The Psychologist Should Attempt to Obtain an Unbiased Estimate of Operational Predictor Validity in the Population in Which It Will Be Used.* Observed validity coefficients are typically not unbiased (Schmidt, Hunter, McKenzie & Muldrow, 1979). Where range restriction operates to bias validity estimates, the appropriate adjustments should be made whenever the information necessary to

do so can be obtained. Adjustments for criterion unreliability should likewise be made whenever an appropriate estimate of criterion reliability can be obtained. Psychologists should give careful attention to ensuring that reliability estimates used are appropriate to this correction in order to avoid under or over-estimating validity. Both unadjusted and adjusted coefficients should be reported. Researchers should be aware that the usual tests of statistical significance are not applicable to coefficients adjusted for restriction of range and/or criterion unreliability. Nevertheless, the adjusted coefficient is generally the best point estimate one can make of the relationship between predictor and criterion. No adjustment of a validity coefficient for *unreliability of a predictor* should be reported unless one clearly notes that the resultant coefficient is theoretical in nature and not operational.

e. *Where Predictors are to be Used in Combination, Researchers Should Give Careful Consideration to 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 utility of valid selection procedures. When nonlinear selection rules are recommended, a clear rationale (e.g., in terms of administrative convenience or reduced testing costs) should be provided. Tests with linear relationships with job performance can be combined for actual 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, and other implications of each choice.

f. *Researchers Should Guard Against Overestimates of Validity Resulting from Capitalization on Chance.* Especially when initial sample size is small, estimates of the validity of a composite battery developed on the basis of a regression equation should be adjusted using the appropriate *shrinkage formula* or be cross-validated on a new sample. It should be noted that 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, most shrinkage formulas are inappropriate and the alternative is cross-validation unless sample sizes are large.

g. *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. In interpreting the results of validity studies, the researcher should take into account the previous relevant research literature as well as the specific study at hand. A history of similar findings in the research literature lends additional credence to the results of individual studies. On the other hand, dissimilar findings should be viewed with caution.

h. *The Researcher Should Ordinarily Make an Assessment of the Practical Value (or Utility) of the Selection Procedure.* There are several approaches to assessing the practical value of selection procedures. In some cases a judgment that a procedure is of significant practical value can be based on the consideration of validity, 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. More sophisticated estimates of the impact of selection tests on the productivity of selectees can typically be obtained by using regression-based equations (Brogden, 1949; Cronbach & Gleser, 1965; Schmidt, Hunter, McKenzie & Muldrow, 1979). Both productivity gains per selectee and total productivity gains due to use of the procedure are relevant in assessing the practical value of selection procedures.

i. *Data Should be Free from Clerical Error.* Key punching, coding and computational work should be checked carefully and thoroughly.

Content-Oriented Strategies

Content-oriented predictor development or choice, if properly conducted, provides evidence that a selection procedure samples job requirements. The following provides guidance for the development of predictors from which valid inferences can be made.

Appropriate development of a selection procedure on the basis of *content* requires developing the procedure to be an appropriate sample of a specified content domain. If a selection procedure is to be used for employment decisions, the relevant content domain is *performance* (or the knowledge, skill, or ability necessary for performance) on the job, in relevant job training, or on specified aspects of either (Lawshe, 1975). A procedure may be a sample of a given 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, whether scores are to be interpreted as measures of achievement or as measures of work behavior. This discussion is limited, however, to situations in which the assessment is evaluated solely in terms of content sampling. It should be noted that content sampling is as useful in the construction and evaluation of criterion measures as it is for selection procedures used for employment decisions.

In content sampling, any inference about the usefulness of a score must be preceded by a set of inferences about the instrument itself based on the method of its construction (Messick, 1975). For that reason, the emphasis of this section and of its title is on the *development* of content-oriented assessment instruments rather than on inferences from scores. Any evaluation of *existing* selection procedures in terms of adequacy of content sampling might follow parallel considerations.

A. *The Job Content to be Sampled Should be Defined.* That definition should be based on an understanding of the job, organization needs, labor markets, and other considerations leading to personnel specifications and relevant to the organization's purposes. The domain need not be inclusive insofar as any larger domain is concerned. By this we mean that it does not have to cover the entire universe of topics covered in a training course or of duties of a particular job. In fact, there may be many domains in the total content universe for any given job. For both what it *does* and *does not* include, a job content domain should be completely defined and thoroughly described.

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 which is quite general; e.g., eliminating material like specific sales prices which may change from month to month and concentrating on content which 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 in view of 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, and no one except the researcher can determine how general a selection procedure should be. The important thing is that the researcher be aware in advance of conditions which may affect the generality decision; and that the generality decision must have a clear rationale based on the specific selection situation at hand, organizational needs, anticipated changes in technology, equipment, and work assignments, and human and economic considerations. This principle also applies in the development of content-oriented criteria for use in a predictive or concurrent criterion-related study. The degree

to which the results of the study can be generalized will depend 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. There are many instances in which domains must be described differently depending on the exact situation. It may even be necessary to assess possible measurement problems in advance of domain definition. Generally, in the case of work samples, the closer a domain is to the totality of the job, 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 test so that every examinee would have the same amount and kind of dirt to remove. In this situation, it would be appropriate to eliminate such cleaning tasks from the test domain. Similarly, seldom used symbols such as the hyphen or question mark appear in different places on different typewriter keyboards; thus, it might be appropriate to limit a typing domain to alpha and numeric characters which are standard on all typewriters. Also, a short course designed to select persons for a longer course should not be based on a domain involving the totality of the longer course, because the advanced lessons require knowledge gained in the beginning lessons. In this situation, the domain should be defined only in terms of lessons which require no prior knowledge. Again, 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(s).* A content domain should ordinarily be defined in terms of tasks, activities or responsibilities or specific abilities, knowledge, or job skills found to be prerequisite to effective behavior in the domain. This means conducting a *job analysis*. This may be a formal investigation, or the pooled judgments of informed persons such as production engineers, job incumbents, their supervisors, or personnel specialists. (See p. 4.) The term "ability" is difficult to define and distinguish from "skill," and it is important to note here that the use of the former term does not imply that content validity is a sufficient justification for the use of abilities or for such characteristics as empathy, dominance, leadership aptitude, and other broad psychological traits. Justification for the measurement and use of such traits must be based on empirical data rather than content sampling alone. It also follows that many procedures developed from general use in a variety of situations are not appropriate samples of a properly defined domain of job content. In particular, general intelligence tests are not appropriately justified by content sampling.

Job requirements assessed by other than formal tests may be established on the basis of content. Requirements for or evaluation of specific prior training, experience, or achievements can be content valid on the basis of the relationship between the content of the training, experience, or achievements and the content of the job for which the training, experience, or achievements are evaluated or required. The critical consideration is the similarity between the products, knowledges, skills, or abilities demonstrated in the experience, training or achievements and the products, knowledges, skills, or abilities required on the job, whether or not there is a close resemblance between the experience, training and achievements as a whole and the job as a whole.

D. *Job Content Domains Should be Defined in Terms of Those Things an Employee is Expected to Do Without Training or Experience on the Job.* It is important to delineate what knowledge, skills, and abilities an employee is expected to have before placement on the job, and define the selection domain in those terms. This definition process often is not simple. There is a fine line between what an employee brings to the job and what he or she is taught on the job. In many instances, those who bring more learning to the job require shorter or different training than others. It is incumbent on the investigator to seek the

appropriate balance between selection and training and define the content domain for the procedure in accordance with this balance (Goldstein, 1980). The point here is that selection does not occur independently of training and this fact must be taken into account. The principle stated here does not preclude relegating different levels of the same ability to selection and training. For example, the fact that an employee is taught to read and interpret company technical manuals does not mean that the job applicant should not be evaluated for basic reading skills.

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. On the other hand, a single activity may be so important that it constitutes a single domain for measurement purposes. For example, a truck driver must be able to drive a truck. The fact that he or she may perform other functions is irrelevant to developing a measure of driving skill or ability.

F. *Sampling of a Job Content Domain Should Ensure that the Measure Includes the Major Elements of the Defined Domain.* Sampling the job content domain is the process of constructing or choosing the selection procedure. If the domain is defined properly; e.g., excludes those things not appropriately measured, learned on the job, or trivial, there should be little difficulty in moving fairly directly from domain elements to selection procedure elements. Any sampling done at this stage should have some rationale; e.g., the most critical elements are chosen. Random sampling is not usually appropriate in this area. Generally, the acceptability of the selection procedure rests on the extent to which elements of the procedure domain match elements of a job content domain.

G. *A Test Developed on the Basis of Content Sampling Should Have Appropriate Measurement Properties.* Linn (1979) has pointed out that there are contradictions between strategies based on domain considerations and those based upon score considerations. A very simple example of these problems is the question of what to do with a test item which is either too easy or too difficult and thereby contributes nothing to the total score variance. Under a score or norm referenced strategy, the item would be eliminated. Using a domain or criterion referenced strategy, the item would be retained.

Although there is much opportunity for further discussion in this area, it appears that for selection purposes, as opposed to achievement measurement purposes, the investigator should resolve many of the differences between the strategies in the direction of norm referenced strategies. For purposes of selection, it is appropriate to consider the instrument involved as predictive in nature in the sense that the evaluation is intended to measure the probability of job success. As can be noted from previous sections, if one considers even limited needs for generality, the selection procedure developed will ordinarily be less than a representative sample of any content domain, although in the development process, every reasonable effort should be made to maintain content domain relevance from the selection procedure. The following suggestions are made to provide effective measurement in a predictive instrument.

1. Where feasible, the selection procedure should be subjected to pretesting and an analysis of the procedure in terms of the means, variances, and intercorrelations of its parts.

a. Parts which do not contribute to the total variance should be considered for elimination. Any replacement parts should reflect the same area of the content domain as those parts which were eliminated.

b. When a critical score is specified in advance and is not expected to fluctuate with labor market conditions or other events, parts which yield maximal discrimination at that score level should be selected. However, any selection of parts should take into consideration the sampling of the original content domain; i.e., a test item from one subject matter area should not normally be replaced with

one from another subject matter area simply on the basis of item statistics. Furthermore, any efforts to increase total variance should take into consideration the need to reflect the content domain.

c. Questions dealing with intercorrelation of parts should be dealt with judiciously. Extreme redundancy of measurement should be avoided. Redundancy reduction may be achieved to some extent through reduction of job analysis data preliminary to domain definition, or it may be effected through analysis of trial administrations of the selection procedure. Redundancy reduction in content-oriented test construction is somewhat analogous to test selection through multiple regression techniques in criterion-related methodology. However, in reducing redundancy one should consider the need for a certain amount of redundancy to provide adequate reliability of measurement. Well constructed parts which do not correlate with other parts or a total score should not necessarily be eliminated. Many domains relative to job performance are multidimensional. For example, a typist who can hit the correct keys cannot necessarily do the arithmetic necessary to do the set-up of the columns for a numerical table. If the lack of correlation among selection procedure parts is merely reflective of 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 when work samples are involved. Equipment may wear or function variably; scoring variations may occur; a desire to minimize testing time may result in taking a sample too small to ensure reliable results; practice and fatigue effects may also be a problem. The foregoing is not meant to suggest that work samples are inappropriate; obviously, for many situations, they are appropriate measuring devices. However, unreliable work sample scores are not to be preferred over well constructed, reliable paper-and-pencil scores.

3. Scoring schemes for content-oriented tests should be ascertained to be correct. Multiple correct answers should be avoided unless they are clearly justified by information about the job.

4. Interpretation of content-oriented selection procedures may reflect the measurement properties of the given procedure. If a selection instrument yields reliable results, and provides adequate discrimination in the score ranges involved, persons may be ranked upon 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 in order. A critical score is also in order in situations such as those in which the greater speed at which a typist can type cannot be reflected in production because of equipment or process limitations. In this case, the selection procedure should be designed with the limiting conditions considered.

H. *Persons Used in Any Aspect of the Development or Choice of Selection Procedures to be Defined on the Basis of Content Sampling Should be Clearly Qualified.* Panels of 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. The investigator should resist accepting people who are not thoroughly technically qualified. Furthermore, any individuals involved in the procedure construction or choice process should be thoroughly trained in those aspects of measurement necessary for their roles.

Generality of Validation Efforts

Only that which is generalizable beyond the specific, immediate situation will have much meaning or practical use except to that specific situation. As was pointed out earlier, the degree of generality to be sought must be determined from the total situation. Many questions regarding generality are still open to debate, but they are a matter of concern regardless of the validation strategy used. The

two topics most closely associated with generality—construct strategies (which provide the ultimate in meaning and generality) and validity generalization are discussed in this section.

A. *The Use of Construct Applications in Employee Selections.* That which has been called "construct validity" in various publications (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978) is really an extension of the traditional concept of construct validity (American Psychological Association, Division of Industrial-Organizational Psychology, 1975). Lerner (1977) has spoken of the traditional (Cronbach & Meehl, 1955) concepts of construct validation as "an ideal perpetually to be sought, not a workable standard which can be legally imposed!" (p. 302-303). The same might well be said for professional advice. Consequently, exact principles for any extension or modification of this concept are difficult to prescribe. Investigators are advised that constructs are essentially theoretical concepts supported by disconfirmatory research. There is need for considerable research to support meaningful interpretations of many selection procedure variables. It appears that at present the best support is in the area of mental abilities (Ekstrom, 1973). The investigator is obligated to search the literature carefully regarding the disconfirmatory research supporting the construct he or she wishes to use in validation. The use of construct definitions without appropriate research support is unacceptable. The investigator is obligated to do his or her own research when the literature does not contain adequate data. Thus, the extension of construct validity often involves considerably more effort than other validation strategies. However, this effort needs to be undertaken and communicated. It is probably only through the generation of more theoretical data in the area of personnel selection that many of the pressing problems facing personnel selection specialists today can be solved. More must be known about the meaning of selection procedure scores (Dunnette & Borman, 1979), so that future research work in this area can go beyond the confines of specific procedures for specific jobs.

Although little guidance is offered here for an investigator faced with a selection situation in which traditional types of validation methodology are inappropriate or infeasible, it should be noted that there is growing concern by a number of researchers about the problems in this area. Cronbach (1980), for example, has proposed strategies less complicated than the traditional construct validation model. Considerable debate will certainly continue to center around validation strategies and the investigator is advised to keep informed and evaluate carefully the literature in this area. In the meantime, those evaluating validation efforts should consider the total evidence relative to the evaluative task and not be constrained by previous conceptions of fixed models of validation.

B. *Validity Generalization.* Classic psychometric teaching has long held that validity is specific to the research study and that inability to generalize is one of the most serious shortcomings of selection psychology (Guion, 1976). As has been pointed out previously, current research is showing that the differential effects of numerous variables may not be as great as heretofore assumed. To these findings are being added theoretical formulations, buttressed by empirical data, which propose that much of the difference in observed outcomes of validation research is due to statistical artifacts (Callender & Osburn, in press; Schmidt, Hunter, Pearlman & Shane, 1979). Continued evidence in this direction should enable further extensions of validity generalization. Cooperative validation efforts being carried on by a number of trade and industry associations will provide the data necessary for evaluation. Such cooperative efforts are to be applauded and encouraged.

Implementation

Validation, discussed in the preceding section, is the investigatory phase in the development or choice of selection procedures. Whatever the outcome of such

research, the researcher should prepare a report of the findings. The importance of documentation in the form of such a report is especially great if the assessment procedure is to be adopted for operational use. Many valid selection programs fail at the point of their implementation. The following principles are intended to assure effective and proper use of measures found valid.

A. *Research Reports and Procedures Manuals.* Validation research is rarely undertaken for the sake of research. Some general guidance on what to do after research follows.

1. Whenever an assessment procedure is made available for use in employment 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 assessment to know what was done, to draw independent conclusions in evaluating the work, and to replicate the study when feasible. This obviously means documentation which covers all essential variables, samples, and treatments. A basic principle in the preparation of such reports is that they should not be misleading. Research findings which might qualify the conclusions or the generalizability of results should be reported.

2. Informational material distributed should be accurate, complete for its purposes, and written in language that is not misleading. 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 others within the organization an impression that an assessment program is more useful than it really is.

3. Research reports and procedures manuals should be reviewed periodically and revised as needed. Any changes in use or in research data that would make any statement in such documents incorrect or misleading should result in revision.

4. Research reports or procedures manuals should help readers make correct interpretations of data and should warn them against common misuses of information.

5. Procedures for administration or other use of a selection procedure should be written by a psychologist or other appropriately trained professional.

6. 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.

7. Any claim made for any selection procedure should be supported in documentation with appropriate research evidence.

8. The procedures manual for persons who administer tests (or use other procedures) should specify the procedures to be followed and emphasize the necessity for standardization of administration, scoring and interpretation. These instructions should be clear enough for all persons concerned to know precisely what they are supposed to do. It should be made clear to everyone involved that failure to follow standardized procedures may render the research results irrelevant to some degree. One must be both insistent and persuasive to get people to understand both the nature of and the need for standardized administration of tests or the use of other procedures. Periodic seminars run by psychologists or 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 nonstandardized procedures may have been used and where the results show no serious impairment of validity. In such situations, the degree of standardization is shown to be relatively unimportant. This should not be assumed without investigation.

9. Any scoring or scaling procedures should be presented in the procedures

manual with as much detail and clarity as possible to reduce clerical errors in scoring and to increase the reliability of any judgments required. When keys must be kept confidential, this material should be made available only to persons who do the actual scoring or scaling of responses.

10. A research report should contain clear and prominent descriptions of the samples used in the research. Such information should also be summarized on any accompanying report forms in which scores are given with normative interpretations such as centiles or expectancies of success.

Ordinarily, norm tables are less useful than expectancy charts for employment decisions. One should recognize, of course, that the expectancy chart is a normative interpretation of test scores; i.e., it indicates the proportion of a specific sample of candidates who reach a specified level of success. Norm tables may be useful in identifying the effects of a cutting score, even if not in interpreting individual employment procedure scores.

11. Any normative reporting should include measures of central tendency and variability and should clearly establish the nature of the normative data given; i.e., centiles, standard scores, expectancies, predicted levels of attainment, etc.

12. Any derived scale used for reporting scores should be carefully described in the research report or procedures manual. Whether using standard derived scores (such as those described in general textbooks on measurement) or "home-grown" scales (such as "qualified," "marginal," or "unqualified"), the researcher should make clear their logical and psychometric foundations.

13. Assumptions of validity generalized from promotional literature or testimonial statements may not be used as evidence of the validity of the procedure. Validity evidence should be built on a foundation of systematic procedures like those discussed in this document.

B. *Use of Research Results.* Application of data in the operational situation must be considered. There are a number of judgments to be made here.

1. It is the responsibility of the researcher to recommend specific methods of score interpretation to the user(s). Although the management of the organization usually retains the final decision on whether to use a specific selection procedure, it is the responsibility of the researcher to make recommendations on this question and on questions of *how* the procedure is to be used. The recommended use should be consistent with the procedures with which validity was established.

2. The utility of a selection procedure should be considered in deciding whether to apply it operationally. In reaching the decision, consideration should be given to relative costs and benefits to both the organization and its employees. It is not recommended that procedures of marginal usefulness be applied, but a procedure with at least some demonstrated utility is ordinarily preferable to one of unknown validity or usefulness. Under usual circumstances, utility has a direct relationship to the coefficient of correlation (Brogden, 1949; Cronbach & Gleser, 1965) and, as mentioned previously, some methods of doing cost-benefit analysis on this basis have been developed (Schmidt, Hunter, McKenzie & Muldrow, 1979).

3. Selection standards 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 scores on these procedures is appropriate. It is not necessary to add any underlying trait assumptions in order to rank. As has been pointed out,

in some circumstances, such as those where a production line limits the speed at which a worker can produce, a *fixed critical* score may be in order.

It is to be pointed out that judgment is necessary in all critical score establishment. A fully dependable numerical basis for a critical score is seldom, if ever, available. The only justification which can be demanded is that critical scores are determined on the basis of a reasonable rationale. This may involve such factors as estimated cost-benefit ratio, 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 decision (i.e., pass-fail points), the rationale or justification should be made known to users. This principle does not recommend critical scores in preference to other interpretive methods. Rather, the point is that, if critical scores are to be established, there should be some rationale and this rationale should be clearly communicated to users.

4. Employers should provide reasonable opportunities for reconsidering candidates whenever alternative forms for assessment exist and reconsideration is technically feasible. Under at least some circumstances, employers should allow candidates to reapply. There might be any of several reasons for questioning the validity of prior assessment for any given person. Where there has been opportunity for new learning, retesting or reevaluating is usually a desirable practice.

5. The use of a predictor, particularly a noncognitive predictor, should be accompanied by systematic procedures for developing additional data for continued research. Changing social, economic, technical, or other factors may operate over time to alter or eliminate validity. Periodic research is therefore necessary. A serious problem is that 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, 1966). There is no well-established technology for checking validity of instruments in use. However, researchers are urged to exercise their ingenuity to observe the principle that validity once demonstrated cannot be assumed to be eternal.

6. All persons within the organization 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 psychologist or other person in charge of any selection program should know measurement principles and the limitations on the validities of interpretations of assessments. That person should understand the literature relevant to the selection procedure use or employment problems. Other persons in the organization may have some responsibilities related to the selection program. It is the responsibility of the person in charge to see to it that such persons have the training necessary to carry out those responsibilities competently. These considerations suggest the need for planned approaches to the training of technicians and managers involved in assessment procedures and in the interpretation of assessments.

7. Researchers should seek to avoid bias in choosing, administering, and interpreting selection procedures. They should try to avoid even the appearance of discriminatory practice. This is another principle difficult to apply. It goes beyond data analysis. The very appearance of bias may interfere with the effective performance of a candidate in the assessment situation. At the very least, a selection procedure user can create an environment that is responsive to the feelings of all candidates, insuring the dignity of all persons.

8. Researchers should recommend procedures which will insure periodic audit of selection procedure use. Departures from established procedures often occur over time. New findings in psychological or psychometric theory, or new social criticisms, may be relevant to one or more of the assessment procedures in use. The principle is that it should not be left to chance to find examples of misuse or of obsolete data. Some systematic plan for review should be followed.

9. The researcher should recommend procedures which will assure clerical

accuracy in scoring, checking, coding, or recording selection procedure results. This principle applies to the researcher and to any agent to whom he or she has delegated responsibility. The responsibility cannot be abrogated by purchasing services from an outside scoring service.

10. The researcher must make considered recommendations for the operational use of a predictor in any instance in which the data appear to indicate differential prediction. A finding of differential prediction should not automatically lead to differences in predictor use for different groups. For example, if the study were based upon an extremely large sample, a finding of statistically significantly differential prediction may have little practical impact. For another example, data apparently indicating differential prediction may be due to statistical artifacts or may suggest courses of action inconsistent with societal goals. In such situations, the reasonable course of action would be to recommend uniform operational use of the predictor for the different groups (or perhaps conduct further research).

Should a finding of differential prediction be compelling enough to warrant other action, possible approaches to dealing with it are (1) replacing the selection procedures involved, or (2) using the selection procedure operationally, taking into account the differences in prediction results. Action under the second alternative should be in accordance with the definition of fairness upon which the study indicating differential prediction was based. In the absence of a compelling finding of differential prediction, the researcher should not recommend differential use of a selection procedure.

11. The researcher or other user is responsible for maintaining security. This means that all reasonable precautions should be taken to safeguard materials and that decision makers should beware of basing decisions on scores obtained from insecure selection procedures.

This principle is difficult to apply to non-test predictors such as judgments reached in an employment interview. Nevertheless, the principle of security as a means for standardization and preservation of validity may be applied to other variables as well. As an illustration of the extension of this principle, reference checks, for example, should be held confidential. Certainly, actual selection procedure scores should be released only to specified persons qualified to interpret them. Every reasonable effort should be made to avoid situations in which injury to the person or damage to the program can result.

12. All implementation procedures should be designed 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, in the operation of the selection program, have advantages, such as coaching, that were not present during the validation effort. Finally, public disclosure of test content should be recognized as a serious threat to the validity, reliability, and subsequent development of testing procedures.

13. In making interpretations of scores, the researcher should be aware of situational variables which may on rare occasion introduce error. An individual score may lead to invalid inferences because of unusual features of the situation (e.g., uncommon distractions), exceptional characteristics of the individual (e.g., a sensory or physical handicap) or the passage of time (e.g., demonstrable new learning since evaluation occurred). Sometimes these may form a basis for re-evaluation. They may suggest the consideration of other information. The principle is that some degree of judgment be retained in the interpretation of scores obtained in circumstances differing from those in the validation research. Perhaps a better statement of the principle is that judgment should not be automatically ruled out in all situations.

14. Any record of scores should be kept in terms of raw scores. There have

been many instances in which data maintained in terms of derived scales⁵ have been found inappropriate for further research.

15. Information should not be available for use in personnel decisions when it may no longer be valid. It is recognized that some traits or characteristics are more stable than others but, as a general principle, it is poor practice to retain test scores or other evaluations in personnel files long after the scores were obtained. Personnel files should be purged of data rendered potentially invalid by new experience, aging, maturation, or other personal change—or by changes in jobs or organizations—so that no one will make inferences on such scores. However, appropriate data should be separately retained for future research.

16. When reporting results, the researcher should consider the level of knowledge of the person receiving the report. The report should be in terms likely to be interpreted correctly by persons at that level of knowledge. Ordinarily, scores should not be reported to candidates or to managerial personnel unless they are explained carefully to make certain that interpretations are correct. In particular, one should not report scores to persons who may be asked later to provide criterion ratings for validation.

17. Scores on many tests developed for educational use are given in derived score form as I.Q.'s or a grade-equivalent context. Such terms are to be avoided. These terms are highly subject to misinterpretation and not likely to be directly meaningful for employment use. 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 to bona fide job candidates. Casual administration of selection procedures to supervisors and others who have no real need to take them can result in breaches of security and, at times, cause personal injury. This principle does not preclude administration for research purposes under appropriately controlled conditions.

Legislation, Regulation, and Court Decision

Opening paragraphs of this document carried a caveat prepared for the 1974 *Standards* and repeated in full in the 1975 *Principles*. One sentence of this caveat reads, "This document is prepared as a technical guide for those within the sponsoring professions; it is not written as law." (See *Statement of Purpose*, p. 4.) Nevertheless, it would be folly for the researcher or practitioner to ignore relevant legislation, subsequent rule-making, and case law in developing strategies for the validation and use of personnel selection procedures (Sparks, 1977).

At the federal level (generally the most important since it typically preempts state legislation if there is conflict) the basic historical statute referring to testing is Title VII of the Civil Rights Act of 1964. This is the basic authority for the various guidelines on employee selection procedures issued by the Equal Employment Opportunity Commission and other EEO enforcement agencies (1978). The U.S. Supreme Court noted initially that the interpretations of the enforcement agency were entitled to "great deference" (*Griggs v. Duke Power Co.*, 1971). Later cases (*Albemarle v. Moody*, 1975; *Washington v. Davis*, 1976) gave further interpretations involving use of selection procedures. Hundreds of lower court decisions have been rendered based on EEOC guidelines and on interpretations of the U.S. Supreme Court decisions (U.S. Office of Personnel Management, 1979; The Psychological Corporation, 1978). These guidelines and the court decisions sometimes conflict with precepts set forth in these *Principles*. More recently, however, the Supreme Court has been reexamining the relationship between agency guidelines and the judgments of psychometric experts as expressed in consensual documents like the APA *Standards* and the Division 14 *Principles*. In some cases such apparent conflicts have been resolved in a manner consonant

with the latter rather than the former (Lerner, 1978). Nevertheless, the researcher or practitioner may need to perform additional analyses in order to satisfy these guidelines or case law.

Recently, a new legislative approach has been taken in the area of testing. Generally referred to as "Truth in Testing" legislation, proposals in the U.S. House of Representatives (Gibbons, 1979; Weiss, 1979) would require (among other things) that test publishers and users make available to examinees copies of their completed test papers or answer sheets with the correct answers marked, completely destroying the security of the tests and creating numerous inimical side effects which would decrease, if not destroy, the validity of the tests. To date, only state legislation has been passed (California and New York). Bills have been introduced in several other states. Researchers and practitioners should be alert to these developments.

Glossary

- Assessment procedure:** any method used to evaluate characteristics of persons.
- Battery:** a combination of two or more scores that predict job performance better than the individual scores alone.
- Bias:** any constant error; any systematic influence on measures or on statistical results irrelevant to the purpose of measurement.
- Coefficient of correlation:** an index number, which may be positive or negative, ranging from 0.00 to 1.00, indicating the extent to which two variables covary.
- Concurrent validity:** a demonstrated relationship between job performance and scores on tests administered to present employees.
- Concurrent validity model:** an approach to validation in which predictor and criterion information are obtained for present employees at approximately the same time.
- Confidence interval:** the bounds on a measurement that define a certain probability that the interval will include the parameter of interest.
- Confidence limits:** the upper and lower limits of the confidence interval.
- Configural scoring:** the assignment of weights to paired variables so that the implication of one predictor score depends upon the level of the second predictor score.
- Construct:** as used here, a trait of individuals inferred from empirical evidence (e.g., numerical ability).
- Construct validity:** a demonstrated relationship between underlying traits or "hypothetical constructs" inferred from behavior and a set of test measures related to those constructs. Construct validity is not established with a single study but only with the understanding that comes from a large body of empirical evidence.
- Contamination:** any systematic influence on measures or on statistical results irrelevant to the purpose of measurement; any bias or error.
- Content domain:** a body of knowledge and/or a set of tasks or other behaviors defined so that given facts or behaviors may be classified as included or excluded.
- Content validity:** a relationship between job performance and a test that is self-evident because the test includes a representative sample of job tasks. (A typing test is content-valid for a stenographer's job.) What constitutes a representative sample of tasks is determined through a job analysis.
- Correlation:** the degree to which two or more sets of measurements vary together; e.g., a positive correlation exists when high values on one scale are associated with high values on another.
- Credibility limits:** a term used in Bayesian statistics, roughly equivalent to confidence limits.
- Criterion:** some measure of job performance, such as productivity, accident rate, absenteeism, reject rate, training score, and so forth. It also includes subjective measures such as supervisory ratings.
- Criterion-related validity:** the statistical statement of the existence of a relationship between scores on a predictor and scores on a criterion measure.
- Critical score:** cutting score; a specified point in a predictor distribution below which candidates are rejected.
- Cross validation:** the application of a scoring system or set of weights empirically derived in one sample to a different sample (drawn from the same population) to investigate the stability of relationships based on the original weights.

- Derived score:** a scale of measurement using a system of standard units (based perhaps on standard deviations or centiles), to which obtained scores on any original scale may be transformed by appropriate numerical manipulation.
- Expectancy table:** a table or chart used for making predictions of levels of criterion performance for specified intervals of predictor scores.
- Feasible:** capable of being done successfully; i.e., in criterion-related research, economically practical and technically possible without misleading or uninterpretable results.
- Job analysis:** a method of analyzing jobs in terms of the tasks performed; the performance standards and training content; and the underlying knowledges, skills, and abilities required.
- Linear combination:** the sum of scores (whether weighted differentially or not) on different assessments to form a single composite score; distinguished from nonlinear combinations in which the different scores may, for example, be multiplied instead of added.
- Moderator variable:** theoretically, a variable which is related to the amount and type of relationship between two other variables.
- Normative:** pertaining to norm groups, i.e., the sample of subjects from which were obtained descriptive statistics (e.g., measure of central tendency, variability, or correlation) or score interpretations (e.g., centiles or expectancies).
- Objective:** verifiable; in measurement, pertaining to scores obtained in a way that minimizes bias or error due to different observers or scorers.
- Operational independence:** gathering of data by methods that are different in procedure or source so that measurement of one variable, such as a criterion, is not influenced by the process of measuring another variable.
- Predictive validity:** a demonstrated relationship between test scores of applicants and some future behavior on the job.
- Predictive validity model:** an approach to validation in which predictor information is obtained at or near the time of hire and criterion information is obtained at a later date.
- Predictor:** a measurable characteristic used to predict criterion performance, e.g., scores on a test, judgments of interviewers, etc.
- Psychometric:** pertaining to the measurement of psychological characteristics such as aptitudes, personality traits, achievement, skill, knowledge, etc.
- Raw score:** the unadjusted score on a test, usually determined by counting the number of correct answers but sometimes determined by subtracting a fraction of the wrong answers from the number of correct answers.
- Regression equation:** an algebraic equation which may be used to predict criterion performance from specific predictor scores.
- Relevance:** the extent to which a criterion measure accurately reflects the relative standing of employees in important job performance dimensions or behaviors.
- Reliable:** consistent or dependable; repeatable; reliability refers to the consistency of measurement.
- Replication:** a repetition of a research study designed to investigate the generality or stability of the results.
- Restriction of range:** a situation, varying in degree, in which the variability of data in a sample is less than the variability in the population from which the sample has been drawn.
- Score:** any specific number in a range of possible values describing the assessment of an individual; a generic term applied for convenience to such diverse kinds of measurement as tests, production counts, absence records, course grades, or ratings.
- Standard deviation:** a statistic used to describe the variability within a set of measurements, based on the differences between individual scores and the mean.

Approved For Release 2001/11/07 : CIA-RDP00-01458R000100110001-9

- Standard score:** a score which describes the location of a person's score within a set of scores in terms of distance from the mean in standard deviation units; may include scores on certain derived scales.
- Suppressor variable:** a predictor variable essentially unrelated to the criterion, but highly related to a second predictor, which presumably reduces the invalid variance in the latter when both are entered into a multiple R.
- Synthetic validation:** an approach to validation in which the validity of a test battery put together for a specific use may be inferred from prior research relating predictors to specified and relevant criterion elements.
- Transformed score:** any raw score that has undergone a transformation in scale (usually linear) such that the transformed scores have a predetermined mean and standard deviation.
- Utility:** the practical usefulness of a relationship (such as a validity coefficient) that allows the user to make better predictions, save money, improve efficiency, and so forth.
- Validation:** the process of investigation (i.e., research) through which the degree of validity of a predictor can be estimated. (Note: laypersons often misinterpret the term as if it implied giving a stamp of approval; they should recognize that the result of the research might be zero validity.)
- Validity:** the degree to which inferences from scores on tests or other assessments are justified or supported by evidence.
- Validity coefficient:** a coefficient of correlation showing the strength of relationship between predictor and criterion.
- Validity generalization:** the transportability of validity evidence; the application of validity evidence obtained in one or more situations to other situations.
- Variability:** the extent of individual differences in a particular variable.
- Variable:** a quantity that may take on any one of a specified set of values.
- Variance:** a measure of variability; the square of the standard deviation.

References

- Albemarle Paper Co. v. Moody*. 422 U.S. 405 (1975), 9 EPD 10230, 10 FEP 1181.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *Report of the Joint AERA, APA, NCME Committee for Review of the Standards for Educational and Psychological Tests*. Unpublished manuscript, 1979. (Available from American Psychological Association, Washington, DC).
- American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. *Standards for educational and psychological tests*. Washington, DC: American Psychological Association, 1974.
- American Psychological Association, Division of Industrial-Organizational Psychology. *Principles for the validation and use of personnel selection procedures*. Dayton, OH: Author, 1975.
- Baehr, M. E. *Skills and attributes inventory*. Chicago, IL: The University of Chicago, Industrial Relations Center, 1971.
- Bemis, S. E. Occupational validity of the general aptitude test battery. *Journal of Applied Psychology*, 1968, 52, 240-244.
- Brogden, H. E. When testing pays off. *Personnel Psychology*, 1949, 2, 171-183.
- Callender, J. C., & Osburn, H. G. Development and testing of a new model of validity generalization. *Journal of Applied Psychology*, in press.
- Christal, R. E., & Weissmuller, J. J. *New Comprehensive Occupational Data Analysis Programs (CODAP) for analyzing task factor information* (AFHRL Interim Professional Paper No. TR-76-3). Lackland Air Force Base, TX: Air Force Human Resources Laboratory, 1976.
- Clarey, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 1968, 5, 115-124.
- Cohen, J. *Statistical power analysis for the behavioral sciences*. New York: Academic Press, 1977.
- Cole, N. S. Bias in selection. *Journal of Educational Measurement*, 1973, 10, 237-255.
- Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology*. Chicago: Rand McNally, 1976, 223-326.
- Cook, T. D., & Campbell, D. T. *Quasi experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally, 1979.
- Cornelius, E. T. III, Carron, T. J., & Collins, A. N. Job analysis models and job classification. *Personnel Psychology*, 1979, 32 (4), 693-708.
- Cronbach, L. J. Equity in selection—where psychometrics and political philosophy meet. *Journal of Educational Measurement*, 1976, 13 (1), 31-41.
- Cronbach, L. J. Selection theory for a political world. *Public Personnel Management*, 1980, 9 (1), 37-50.
- Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions* (2nd edition). Urbana, IL: University of Illinois Press, 1965.
- Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, 52, 281-302.
- Darlington, R. B. Another look at "cultural fairness." *Journal of Educational Measurement*, 1971, 8, 71-82.

- Dunnette, M. D. Personnel selection and job placement of disadvantaged and minority persons: Problems, issues, and suggestions. In H. L. Fromkin & J. J. Sherwood (Eds.), *Integrating the organizations*. New York: Free Press, 1974, 55-74.
- Dunnette, M. D., & Borman, W. C. Personnel selection and classification systems. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual Review of Psychology* (Volume 30). Palo Alto, CA: Annual Reviews Inc., 1979, 477-525.
- Einhorn, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, 1971, 75, 261-269.
- Ekstrom, R. E. *Cognitive factors; some recent literature* (Technical Report No. 2, ONR Contract N00014-71-C-0117NR150-329). Washington, DC: Office of Naval Research, July 1973.
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. Adoption by four agencies of uniform guidelines on employee selection procedures (1978). *Federal Register*, 1978, 43, 38290-38315.
- Fine, S. A., & Wiley, W. W. *An introduction to functional job analysis: Methods for manpower analysis* (Monograph No. 4). Kalamazoo, MI: W. E. Upjohn Institute for Employment Research, 1971.
- Gibbons, S. *Truth in testing act of 1979*. H. R. 3564, 96th Congress, 1st session, April 10, 1979.
- Goldstein, I. L. Training in work organizations. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual Review of Psychology* (Volume 31). Palo Alto, CA: Annual Reviews, Inc., 1980, 229-272.
- Griggs v. Duke Power Co.* 401 U.S. 424 (1971), 3 EPD P8137, 3 FEP 175.
- Guion, R. M. Synthetic validity in a small company: A demonstration. *Personnel Psychology*, 1965, 18, 49-63.
- Guion, R. M. Employment tests and discriminatory hiring. *Industrial Relations*, 1966, 5, 20-37.
- Guion, R. M. Recruiting, selection and job placement. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology*. Chicago: Rand McNally, 1976, 777-828.
- Lawshe, C. H. A quantitative approach to content validity. *Personnel Psychology*, 1975, 28 (4), 563-575.
- Lerner, B. *Washington v. Davis: Quantity, quality, and equality in employment testing*. In P. Kurland (Ed.), *The Supreme Court Review* (1976 vol.). Chicago: University of Chicago Press, 1977.
- Lerner, B. The Supreme Court and the APA, AERA, NCME test standards: Past references and future possibilities. *American Psychologist*, 1978, 33, 915-919.
- Linn, R. L. Fair test use in selection. *Review of Educational Research*, 1973, 43, 139-161.
- Linn, R. L. Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology*, 1978, 63, 507-512.
- Linn, R. L. *Critical issues in construct validity*. Paper presented at Educational Testing Service Construct Validity Colloquium, Princeton, NJ, October, 1979.
- McCormick, E. J. *Job analysis: Methods and applications*. New York: AMACOM, 1979.
- McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology*, 1972, 56, 347-368.
- McCormick, E. J., & Mecham, R. G. Job analysis data as a basis for synthetic test validity. *Psychology Annual*, 1970, 4, 30-35.
- Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, 30 (10), 955-966.

- Novick, M. R., & Jackson, P. H. *Statistical methods for educational and psychological research*. New York: McGraw-Hill 1974.
- Pass, J. J., & Cunningham, J. W. *Occupational clusters based on systematically derived work dimensions* (Final Report). Raleigh, NC: North Carolina State University, Center for Occupational Education, 1977.
- Pearlman, K. Job families: A review and discussion of their implications for personnel selection. *Psychological Bulletin*, 1980, 87 (1), 1-28.
- Pearlman, K., Schmidt, F. L., & Hunter, J. E. Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, in press.
- Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 1976, 13, 3-39.
- Peterson, D. A., & Wallace, S. R. Validation and revision of a test in use. *Journal of Applied Psychology*, 1966, 50, 13-17.
- Prien, E. P., & Ronan, W. W. Job analysis: A review of research findings. *Personnel Psychology*, 1971, 24 (3), 371-396.
- Primoff, E. S. *Summary of job-element principles: Preparing a job-element standard*. Washington, DC: U.S. Civil Service Commission, Personnel Measurement and Development Center, 1971.
- Primoff, E. S. *The J-coefficient procedure*. Washington, DC: U.S. Civil Service Commission, Personnel Measurement and Development Center, 1972.
- Schmidt, F. L., Hunter, J. E., McKenzie, R., & Muldrow, T. The impact of valid selection procedures on workforce productivity. *Journal of Applied Psychology*, 1979, 64 (6), 609-626.
- Schmidt, F. L., Hunter, J. E., & Pearlman, K. Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology*, in press.
- Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. *Personnel Psychology*, 1979, 32 (2), 257-281.
- Schmidt, F. L., Hunter, J. E., & Urry, V. W. Statistical power in criterion-related validity studies. *Journal of Applied Psychology*, 1976, 61, 473-485.
- Schwartz, D. J. A probabilistic approach to adverse effect, job relatedness and criterion differences. *Public Personnel Management*, 1978, 7 (6), 368-377.
- Sparks, C. P. Guidance and guidelines. *The Industrial-Organizational Psychologist*, 1977, 14 (3), 30-33.
- Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd Ed.). Washington, DC: American Council on Education, 1971.
- The Psychological Corporation. *Summaries of court decisions on employment testing: 1968-1977*. New York: Author, 1978.
- Thorndike, R. L. Concepts of culture-fairness. *Journal of Educational Measurement*, 1971, 8, 63-70.
- U.S. Office of Personnel Management. *EEO Court Cases* (September 1979 Revision). Washington, DC: Author, 1979.
- Wallace, S. R. Criteria for what? *American Psychologist*, 1965, 20, 411-417.
- Washington v. Davis*. 426 U.S. 229 (1976), 12 FEP 1415.
- Weiss, J. *Educational testing act of 1979*. H. R. 4949, 96th Congress, 1st session. July 24, 1979.