33.

Selection

HAROLD P. BECHTOLDT
The State University of Iowa

The development of techniques and procedures to be used in selecting individuals for various jobs, training opportunities, or forms of therapy has become a major activity for many psychologists, and has resulted in an extensive body of knowledge, a multitude of techniques, and a few generalizations of theoretical value.

The objective of this discussion is to summarize the current work in this field and to present some of the general problems and theoretical considerations. No attempt will be made to list the various devices used in selection, to review the literature, or to treat the standardization of tests, the preparation of norms, or the office procedures required for the proper protection and utilization of classification materials. Statistical procedures will be discussed, but the formulas and computing procedures have been omitted.

Definition of Selection

Selection is defined by Warren (1934) as "the picking out or emergence of a character, an object, or a phenomenon from a group of alternatives in accordance with some standard or principle," and vocational selection is presented as "the process of choosing from a group of applicants for a vocation those most likely to succeed in that vocation."

Four characteristics of selection may be distinguished.

First, selection of a few individuals from among many is a *process*. This process may be applied only once, as in the selection of a given number of punch-press operators for an individual plant, or it may be a recurring process, as in the periodic grading and elimination of graduate students. The process either may be formalized, with definitely stated critical levels of performance, or it may appear only as an attitude indicative of continuous scrutiny of an individual's behavior.

Second, the selection process always involves selection for some purpose: for speed of response, for capacity to lead men, or for any of the nearly limitless activities of human beings.

Third, selection requires procedures for assigning individuals into classes appropriate to the objective of the selection process. The development of these procedures constitutes the *classification problem*. The classifications may be made in terms of any *relevant* attribute of the individual. The procedures utilized may require (1) the obtaining of responses, verbal or nonverbal, from an individual, (2) reports from observers or "judges," (3) records of past performances, or (4) direct observation by the experimenter.

Fourth, selection implies prediction. The goal of a selection process is the prediction of an individual's "success" or his behavior at some future time. Success, of course, must be defined in terms of attributes other

The writer gratefully acknowledges the extensive assistance and constructive criticisms given by Dewey B. Stuit, State University of Iowa.

than those used in the "sorting" operations of the selection process. The forecasting of future behavior constitutes the *prediction problem.*

Selection typically involves two independent assessments separated by a temporal interval. On the basis of the first assessment we set up "predictor categories." The second assessment leads to "criterion categories." The problem, then, is to predict the criterion categories. The predictions represent *estimates* of the *expected values,* and the estimates are determined from empirical investigations of the relations between the two sets of categories. From these investigations we determine the *most probable* criterion categories corresponding to the various predictor categories.

Because the behavior of the individual is continuously modified by external events, predictions of future behavior are usually made conditional upon the occurrence of a specific state of affairs. When predictions are hedged in this manner, they are said to be conditional predictions (Horst, 1941). For example, predictions of parole success may be conditional upon whether or not the parolees return to their previous environment. The isolation of the conditional factors that are relevant to the criterion classifications is a crucial problem in prediction. The relations between such factors and the "success" categories are analogous to the stimulus-response "laws" sought in experimental studies.

The criterion category is a classification based on a performance characteristic. In contrast, the predictor category may represent *any* attribute of the individual, defined by how he appears to others, what he can do, or what he has done. Either of the two classifications (predictor or criterion) may be expressed in terms of a single characteristic or in terms of a set of measures, as in a multiple regression problem. However, the criterion attributes are usually combined to provide a single composite "measure of success" in terms of which the prediction statements are tested. When the selection process is a recurrent one, the criterion classification at an intermediate stage may be considered a predictor category for a later stage, and the final measure of success is then designated the "ultimate" criterion.

SELECTION AS AN ASPECT OF PERSONNEL PSYCHOLOGY

Unfortunately, there are few occasions in personnel selection when ready-made techniques can be applied. The personnel psychologist must meet most situations armed only with a method of approach. He must first study the problem in terms of the purpose of the selection, then formulate working hypotheses, devise the necessary procedures, try them out experimentally, check on the validity of his hypotheses, and finally revise his procedures in accordance with his findings. Each of these steps is a separate problem to be solved before the actual selection of individuals is undertaken.

The psychological analysis of behavior has not progressed to the point where direct application of general laws to the "engineering problems" of selection can be made. Each new problem has a few elements in common with old ones, plus a number of novel aspects, the effects of which can rarely be accurately predicted. Hence there is a demand for continued patient research in the applied field — a demand that usually runs counter to the demands for immediate results. This conflict is probably responsible for a concentration of effort on formalized tests of aptitude and achievement, and for the use of linear prediction equations.

Individual differences, which are often assumed to be a mere source of error in other problems, are the meat of the selection business. If individuals are homogeneous with respect to the predictor attributes, no *differential* classification in terms of these attributes can be made. If individuals are homogeneous with respect to the criterion categories, there is no selection problem.

Individual prediction, it should be pointed out, is to some extent a misnomer. Predictions are properly applied to classes of individuals homogeneous (within limits) with respect to a set of characteristics (Sarbin, 1944). The personnel psychologist is concerned with assigning individuals to classes and with making statements about the expected performances of members of the classes.

Similarity of selection research to other psychological research. Between selection and other areas of psychological endeavor there is a similar objective and a similar point of view. The objective is the formulation of empirical rules based upon observed consistencies in the attributes of individuals, and in the effects upon behavior of various stimulus conditions. The point of view is a behavioristic one in the sense that observations of overt behavior, verbal and nonverbal, and discriminable attributes of individuals are the terms related.

In pushing toward this objective of discovering relations among variables, personnel psychologists are tending to consider their problems strictly in empirical terms. This extreme empiricism is largely a reflection of the specific problems formulated in selection studies, and a reaction against the unwarranted generalizations frequently used in the past. The typical selection project has some specific purpose, and its success is evaluated strictly in terms of the accuracy of its predictions. The empirical relations between the criterion and predictor categories are the pertinent observations; no abstract or theoretical constructs need appear in the expression of these relations. If the procedures work, they are accepted. This justification of selection techniques by a pragmatic criterion may explain Guthrie's characterization of the test movement in applied psychology as ". . . highly useful and practical work, but it has not contributed to psychological theory" (1946).

This is not to suggest, however, that the applied psychologist can make no theoretical contributions and achieve no important generalizations. He appeals to psychological principles when he chooses predictor variables and estimates the effects of situational factors. The agreement between the obtained and predicted observations constitutes a partial test of these principles. Furthermore the applied psychologist accumulates observations on the empirical regularities in human behavior. These regularities may form the basis for useful hypotheses and generalizations.

Both trait concepts and descriptions of stimulus conditions are used by the personnel psychologist, but trait concepts are probably given the greater emphasis in selection because of the difficulty of manipulating environmental conditions.

Trait concepts develop from the obvious fact of conspicuous individual differences. "Few jobs exist which can be performed equally well by practically all of the participants" (Flanagan, 1948). Although some differences among individuals are subject to change, other differences are quite stable. Furthermore, a high degree of specificity is found in the traits associated with the performance of various tasks. This specificity in performance is not explicable in terms of some unreliability of evaluation, for a high degree of consistency within each single area of activity can often be demonstrated.

Consistency of performance in a given domain and independence of performances in different areas have given rise to the notion of functional unities, i.e. traits expressed as abilities, motives, interests, and temperamental variables. The provision for more adequate operational definitions of these postulated traits is one of the outcomes of the factorial approach. It has been suggested that the number of traits necessary to account for the various occupational performance measures may be relatively large (Flanagan, 1947*a*). It should perhaps be noted also that the factorial methods do not always sustain the initial hypotheses concerning the characteristics of a given trait; as often as not these hypotheses are subject to drastic modification in the light of the intercorrelations obtained among the tasks selected to represent the trait (Thurstone, 1947).

The definition of these traits by a set of test situations involves no assumptions about the origins of the traits; their source may lie in the social or in the physiological structure. The traits may be acquired or inherited, easily modifiable or resistant to change, specific to a given situation or common to many situations. Traits having the greatest stability and generality are of the greatest interest, of course, and the personnel psychologist seeks to isolate and define them and to establish relations among them.

Steps in Selection Research

The preceding discussion has emphasized two operations: the classification of behavior and the prediction of behavior. Before these operations can be carried out it is necessary that certain basic research be completed. The following five steps, modified from those listed by Horst (1941) and Sarbin (1944), represent the tasks involved in selection research:

1. The establishment of the criterion categories of "success," which involves the definition of the behavior to be predicted and the development of procedures for the classification of performance.

2. The selection of the attributes on which prediction is to be based, and the establishment of the several predictor categories for each of the attributes (for a representative sample of the population). These prediction attributes are those that are expected to have significant relations with the criterion attributes.

3. The determination of the relations between the criterion categories and the several predictor categories. These empirical relations are then used to predict the criterion category for each individual.

4. The verification of the relations determined on the basis of the original sample by the application of the classification procedures to a new sample of the population. This verification, or cross-validation, constitutes a crucial step that is too often omitted.

5. The application of the selection procedures in the routine situations for which they were developed, provided the stability of the prediction in the cross-validation step has been sustained.

Simplifying assumptions. In carrying out the selection process, we make certain assumptions for the purpose of simplifying the operational procedures. The first of these is that *human attributes that differ in magnitude can be represented by numbers.* Certainly numerical operations are used in the statistical procedures of selection, but are these statistics justified? Two replies have been given to this question. The first answer is based upon the logical implications of measurement in science: that the representation of empirical magnitudes by numbers requires that certain empirical operations be performed (see Chapter 1). "Additive" (or ratio) scales exist for some of the physical properties of responses but not for the more ubiquitous aspects known as qualitative attributes because the operations for the determination of zero and for the process of addition are not available (Bergmann and Spence, 1944; Richardson, 1941*b*). Psychological scales that are more than simple rank-order scales (Gulliksen, 1946; Stevens, 1946) can be constructed for some psychological properties, but they are rarely used in selection problems.

The second answer is that the use of numerals to represent the discriminal differences in attributes serves a practical purpose. Qualitative dichotomous attributes are often assigned numbers, as 0 and 1, and various statistics are computed. Multiple-category qualitative variables representing intensive dimensions are "scaled" in various ways, and numerical scores are determined.

Since the results are useful, these practices will continue. And the continued search for operations that will meet the logical requirements of measurement may eventually place these procedures on a sound basis.

The second of the simplifying assumptions is that *attributes differing in magnitude can be considered as continuous variables.* Classification categories are considered to represent intervals along a continuum. Of course, there is no assurance that the phenomena under consideration are continuous and not discrete. For example, only completed responses on a test are scored. Such responses are usually discussed, however, as though any given response might be represented by some value between zero and one.

The third simplifying assumption current in selection work is *that the criterion magnitudes can be expressed as linear functions of the predictor variables.* Horst (1941) has pointed out that increased accuracy of prediction may require the use of nonlinear relations. Such functions are current in the formulations of other psychological problems, for example, in the various mathematically stated learning theories. However, for a variety of monotonic functions within the range of values conventionally used, a linear equation provides a useful first approximation. For selection purposes, nonlinearity of relations is not crucial since linearity can often be secured by a modification of the measuring devices or by a transformation of the scales or scores. Furthermore, empirical evidence indicates that sampling fluctuations of regression parameters are of such magnitude that very large representative samples would be required to provide accurate estimates of nonlinearity (Flanagan, 1948).

Necessary Characteristics of Classification Procedures

An acceptable categorizing procedure is one that provides stable differentiations between individuals with respect to a characteristic of performance. Differentiations with respect to a single variable are desired, but for many selection purposes complex "multidimensional" sets of attributes are useful. In every case, however, the classification procedures are expected to show *discrimination* and *reliability* (Adkins, 1947; Thorndike, 1949).

Discrimination is the aspect of a classification procedure that is reflected in the number of categories to which significant proportions of individuals are assigned. A procedure that assigns all individuals to a single class fails to discriminate. If each individual is assigned to a separate class, discrimination is maximal. Most classification procedures exhibit a discrimination intermediate between these two extremes.

An index of discrimination has been developed in terms of the obtained number of differentiations between *pairs* of individuals (Ferguson, 1949). The number of possible differentiations is the number of combinations, $n(n-1)/2$. The index δ, derived in terms of s different *possible* score categories, ranges from zero for the test on which everyone has the same score to unity for the test providing a rectangular distribution extending over the entire range of scores. The formulation is

$$\delta = \frac{s}{s-1}\left(1 - \sum_{j=1}^{s} p_j^2\right)$$

where s = the number of possible score categories, and p_j = the proportion of individuals in the jth category, j ranging from 1 to s.

The effectiveness of the classification procedure can also be considered in terms of the possible number of differentiations, i.e. when each individual is assigned to a separate class. In this case, an index δ' is provided by

$$\delta' = \frac{n}{n-1}\left(1 - \sum_{j=1}^{s} p_j^2\right)$$

where n = number of individuals in the sample, and the other symbols have the same meaning as before. A value of unity for δ' is obtained only when $s = n$ and

$p_j = 1/n$. The index δ' expresses the proportion of the total possible number of discriminations that is provided by the classification procedure.

If the number of discriminations made by a classification procedure is used as a criterion of "goodness of discrimination," the best procedure is one that provides a rectangular distribution of scores (or categorizations) for the group. Since a single item, scored correct or incorrect, can be considered as a "test," the most discriminating item ($\delta = 1.00$) is one that has 50 per cent of the cases in each of the two categories (Adkins, 1947). For sets of two or more items, however, the maximum number of discriminations made by the s total test scores (each item scored 0 or 1) may or may not be obtained with items at the 50-per-cent level of difficulty. The intercorrelations of the items and their reliabilities, as well as the difficulty level of each item, must be considered (Brogden, 1946*b*; Carroll, 1945; Flanagan, 1939*b*; Gulliksen, 1945; Richardson, 1941*a*). It has been shown that the variance of the total score (the score defined as the sum of unit-weighted items scored 0 or 1) varies directly with the average intercorrelation of the items and with the average item variance, and varies inversely with the variance of the distribution of item difficulty values. Since a large variance is associated with a flat distribution of test scores, the relations that tend to maximize the test variance will also tend to increase the index of discrimination of the test. A decrease in the average interitem correlation for a given difficulty distribution will lead to a more nearly unimodal distribution of scores and to a smaller test-score variance and fewer discriminations. When the maximum discrimination is desired, a wide range of difficulty values is recommended in order to secure an approximately rectangular distribution of scores (Adkins, 1947; Loevinger, 1947).

If the function of a test is to measure some hypothetical trait or "ability" common to its several items, the difficulty level of the test providing the maximum discrimination will vary with the ability of the subjects. Easy items ($\bar{p} \geq 0.70$) will give the greater number of discriminations for subjects having low ability, and difficult items ($\bar{p} \leq 0.30$) will be more appropriate for subjects having high ability (Adkins, 1947; Richardson, 1936*a*).

The fineness of discrimination required may, however, vary from time to time, depending primarily upon the purpose to be accomplished. If the problem is simply one of accepting or rejecting each individual on the basis of some requirement, only a dichotomous classification is needed. In this case the test-score distribution would be bimodal or U-shaped. When fine discriminations are desired only among individuals at either extreme of the range of ability, the test-score distributions can be definitely skewed or J-shaped, with the flat section of the distribution occurring in the region where the differentiations are to be made. Rectangular distributions are the goal when the number of classes desired is limited only by the errors of measurement and by the range of the sample.

It should be emphasized that this concept of discrimination does *not* refer to the relation between the assessments of two independent classification procedures. Such relations are used to provide an indication of the "validity" of one of the classifications. A classification procedure may discriminate without being valid (for a given criterion).

Reliability is the consistency or stability of the evaluations obtained from repeated observations. Repetitions of a reliable procedure lead to similar scores. Either the score (or classification) can be shown to be "essentially" the same on repeated trials, or the position of the individual relative to the group remains "relatively" constant. The terms "essentially" and "relatively" indicate that some variation is always expected. The nonsystematic or "chance" variations in repeated observations are termed "errors of measurement." If only a small fraction of the total variance of the scores is associated with these errors, the test is reliable; if the fraction is large, the test is less reliable.

Two common indices of reliability are used in selection studies, the *standard error of measurement* and the *reliability coefficient.* The square of the standard error of measurement is an estimate of the variance of the repeated measurements (error variance) of the same individual's performance. The reliability coefficient is defined as one minus the ratio of the average error variance (estimated from the entire set of observations) to the total variance of the test. This coefficient, ranging between zero and unity, is an index of the stability of a classification relative to the classifications of a group of individuals. The standard error of measurement, at least for a given score or classification, is independent of the range of performance in the sample of individuals (Adkins, 1947; Cronbach, 1947; Jackson and Ferguson, 1941; Thorndike, 1949).

The basic assumption made in estimating the reliability of a procedure is that the repeated observations can be considered as measures of the "same thing." The behavior observed is assumed to remain constant; variations are attributed to extraneous factors that operate in a nonsystematic fashion. The definition of the sources of error variance, the development of experimental procedures for securing the relevant data, and the formulation of statistical techniques for the evaluation of the error variance are three general problems of test reliability.

Since the variance included as "error" by the psychologist may arise from many sources, there are several different reliability concepts. They concern repeated measurements in situations involving (1) samples of the individual's behavior, (2) samples of the tasks or operations, and (3) samples of the performance of observers or scorers evaluating a given behavior or attribute. Estimates of the stability of classifications (relative to the variability of the group), over a sample of similar tasks, over a period of time, or over both, are provided by three coefficients: (1) *equivalence,* (2) *stability,* and (3) *stability and equivalence* (Cronbach, 1947).

Three different experimental procedures are used to obtain the data for these three coefficients. The coefficient of *equivalence* is estimated from the data obtained with a single sample of items presented at one time to a group of individuals. The coefficient of *stability* is determined from the results of two or more administrations of the same test at different times. The coefficient of *stability and equivalence* is estimated from two or more samples of items (parallel forms representing samples from the same universe of items) presented at various intervals of time. As the time interval decreases, the coefficient of stability and equivalence for two or more tests approaches the coefficient of equivalence for a single measurement consisting of the combined samples of items.

The statistical operations used to compute these various coefficients stem from the assumption that the error factors are independent of the systematic factors, or that the responses to any item on any trial are independent of the responses to other items on other trials. The mathematical details are presented by various writers (Adkins, 1947; Cronbach, 1947; Guttman, 1945, 1946; Horst, 1949; Hoyt, 1941; Kuder and Richardson, 1937; Jackson and Ferguson, 1941; Loevinger, 1947; Rulon, 1939).

The pertinent sources of variance can be classified in terms of the duration and generality of their influence. These factors may operate in a *persistent* or in a *temporary* fashion (duration) and may be *specific* to one task or *common* to two or more tasks (generality). All reliability formulations consider the systematic nonerror (true) variance to include common persistent factors such as general skills and abilities. Such factors as luck in guessing or the influence of momentary stimulus conditions are assigned to the error variance. The three formulations of reliability, however, assign different combinations of the remaining factors to the error variance.

The *coefficient of equivalence* assigns all the specific-factor variance, either temporary or persistent, to the error variance. The specific factors include all characteristics of the individual (skills or information) or of the task (form of item or terminology) that may cause the performance on *one* item alone to be more or less satisfactory. Since the data are obtained in a single test administration of the sample of items, the true variance will include any common factors that persist over the time required for the test. The Kuder-Richardson (1937) formulation of the reliability concept (and its variants) provides an exact estimate of the coefficient of equivalence if the specified assumptions are met. If the assumptions are not sustained, the Kuder-Richardson formulas provide conservative estimates of this coefficient. The coefficient of equivalence is often determined by splitting the test into halves (as odd versus even items). The estimate obtained from the Spearman-Brown formula in this case may provide either an overestimate or an underestimate of the coefficient, depending upon the comparability of the halves (Kuder and Richardson, 1937). All these estimates assume that all items of the test have been attempted, and they are not applicable to time-limit testing procedures.

This general equivalence formulation of the problem of the consistency of responses to a given set of items has been termed the *internal consistency hypothesis* and is utilized in the factorial methods, in the principal-axes solutions for combining measures, in the definitions of homogeneous tests, as well as in the estimation of a reliability coefficient (Burt, 1936; Horst, 1941; Loevinger, 1947; Wherry and Gaylord, 1943). The concept of a homogeneous test as developed by Loevinger (1947) provides an alternative formulation of the internal consistency of the responses to a single sample of items ordered with respect to the proportion of correct responses to the items. An index of homogeneity is then provided by the ratio of the obtained interitem covariances to the maximum interitem covariances for an ordered arrangement of the items. For both the Kuder-Richardson and the Loevinger formulations the concept of the total variance of the scores, as a function of the item variances and the interitem correlations, is important; these functions also provide an indication of the close relation between discrimination and reliability.

The error variance for the *coefficient of stability* includes all temporary sources of variance; these are the common and specific factors that do not persist from one test administration to another. The temporary factors may include the individual's degree of motivation, health, or degree of skill in the mechanics of taking the test. An estimate of this coefficient is provided by the correlation between repeated presentations of the same test. The estimates tend to vary with the time interval between test administrations. Guttman's (1945, 1946) formulations are based on this concept, although he provides estimates of the lower limits of this reliability coefficient from a single trial. Two of his estimates are comparable to those secured from the more precise of the several Kuder-Richardson formulas.

All temporary factors, common and specific, as well as the persistent specific factors, are included as error variance in the *coefficient of stability and equivalence.* This coefficient is estimated from the correlations between two (or preferably three) parallel or comparable tasks separated by a time interval. The coefficient is affected by the length of the time interval as well as by the degree of comparability of the forms of the test. This coefficient has been considered most appropriate for evaluating the reliability of the classification procedures used for selection purposes.

Basic Concepts of Prediction

In their prediction procedures, psychologists have made extensive use of the concepts of *probability* and of *validity*. The *relative-frequency* concept of probability provides the basis for the mathematical procedures used to forecast future behavior and to evaluate these forecasts.

Probability. A degree of probability attaches to any estimate of future behavior. The best estimate is the one yielding the least error, i.e. the most probable values, for the population as a whole (Bridgman, 1932; Guttman, 1941). For qualitative data these estimates are the criterion categories having modal frequencies at each predictor category in the bivariate or multivariate distribution. For quantitative data the most probable criterion values are the arithmetic means of the criterion variables determined for each predictor interval (Guttman, 1941). These mean values are computed from the line of regression of the criterion on the predictor variable. When the linearity assumption is accepted, these means are determined from the usual single or multiple regression equations. When the criterion represents a dichotomized variable (scored 0 or 1), the linear prediction equation (for quantitative variables) is Fisher's discriminant function (Garrett, 1943; Travers, 1939).

Because predictions are always inexact, an estimate of the error of prediction is required. This takes the form of a probability statement expressing a relation between a proposition and a set of data (Jeffreys, 1937; Sarbin, 1944). The data are the criterion classes into which the individuals fall; the proposition is the statement that such and such criterion classes are associated with each of the various predictor measures. The relative frequencies of the *successes* of these *propositions* then provide an index of the accuracy of the prediction.

These probability statements properly refer to predictions made about members of classes. A class may be defined, for example, by the consistent behavior of an individual in a given situation over a period of time or by the similar behaviors of the individuals in a group (Sarbin, 1944). The statement that the probability of John Brown's success is 80 per cent is inappropriate; the probability properly refers to the accuracy of our predictions about individuals in the same class as John Brown, say Class X, rather than to John Brown himself. Verification of probability statements is possible only in terms of relative frequencies. If additional *relevant* data are available, a more useful designation of John Brown may be as a member of Class Y rather than X, where the relative frequency of successes may be different for the two classes. A class designation for an individual can be modified by extending or reducing the attributes used to define the class. Augmented class definitions are of value if the additional data are relevant to the criterion measures (Sarbin, 1944).

Validity. The discussions of validity by Mosier (1947), Rulon (1947), and Thorndike (1949), among others, have differentiated among several different concepts of validity encountered in selection.

The first of these has been termed "validity by definition" (Mosier, 1947). It requires a judgment concerning the pertinence and comprehensiveness of the operations used to define the characteristic to be measured (Bechtoldt, 1947; Rulon, 1947). This concept of validity involves the *acceptance* of a set of operations as an adequate definition of whatever is to be measured. The concept appears in two logically similar phases of selection research, namely, in the choice of an ultimate criterion measure and in the development of tests for a specified trait or quality of performance. If the criterion is defined as the ability to use a lathe to machine a steel screw within a given period of time, a work sample involving exactly this task would be an "obviously" *valid* measure of the skill, although it might not necessarily be a reliable one. A valid test of skill in arithmetic could be defined, likewise, as one requiring the execution of addition, subtraction, multiplication, and division. In selection studies that use criteria such as these, agreement among competent observers constitutes a measure of the validity of the criterion.

The second type of validity concerns the agreement between evaluations of the same individuals by two nonequivalent measures, one of which is termed the criterion (dependent) variable and the other the predictor measure. The basic similarity between the statistical concepts of validity and reliability is evident in this formulation; the difference between them rests with whether the measures are nonequivalent or equivalent.

Obviously the statistical concept of validity, since it refers to a relation between a criterion measure and some other assessment, is dependent upon the particular criterion used. For each different objective or purpose a different criterion is required. The statistical validity of a measure varies then from one activity to the next. In this sense a test has no intrinsic validity; it has as many validities as there are criterion measures to be predicted.

Validity by assumption, involving "common-sense" judgments about the abilities measured by the test, is both common and dangerous in selection (Mosier, 1947). The blithe disregard of empirical evidence of validity in favor of the appearance, the title, or the reported factorial composition of the test and of the criterion is seen too frequently. Tests entitled "mechanical ability," that have proved useful in selecting lathe operators, may, for example, be applied to the selection of engineering students under the *assumption* that they are valid for that purpose also, but this assumption must be verified.

Still another type of validity has been termed "palatability," or "face validity" (when the term does not refer to validity by definition) (Mosier, 1947). Since this concept concerns the way in which the respondents react to the appearance of the test, it is not an important phase of validity, as the concept is used here.

Validity coefficients. When two variables are continuous and linearly related, the product-moment correlation coefficient is the index of validity most frequently used. The multiple correlation coefficient is used for the index of validity of sets of predictor variables under similar conditions. When the relations are to be determined between attributes and between both qualitative and quantitative variables, agreement as to the appropriate statistic is not so widespread.

When there are artificial or true dichotomies in the criterion and predictor variables, the point biserial, the biserial, the *phi,* the *phi* biserial, or the tetrachoric correlation coefficients are used. The point biserial and *phi* coefficients represent the product-moment correlations for true dichotomies (scored zero and one) for the cases of one and of two dichotomous measures, respectively. The biserial and tetrachoric coefficients represent estimates of the product-moment correlations under the assumptions of continuous and normal distributions and linear relations for the cases of one and of two artificial dichotomous variables, respectively. The *phi* biserial coefficient is derived for the case of one artificial and one true dichotomy scored 0 and 1, using the assumptions required for the biserial coefficient (Thorndike, 1949).

The point biserial and *phi* coefficients are regarded as the appropriate statistics for the determination of regression coefficients, in place of the biserial and tetrachoric coefficients, even for artificial dichotomies, since these coefficients indicate what the effective relation is for the case of linear prediction (Wherry and Taylor, 1946). The magnitude of the point biserial coefficient, however, is a function of the proportions in each part of the dichotomy (e.g. the difficulty level of a test at the point of division). This characteristic has led some workers to prefer the biserial, tetrachoric, and *phi* biserial coefficients for those cases involving artificial dichotomies (Burt, 1944; Thorndike, 1949). If the analysis is concerned with the homogeneity or factorial complexity of a set of observations, the dependence of the point biserial and *phi* coefficients upon the difficulty level may lead to the appearance of "difficulty" factors if the measures are not homogeneous with respect to proportion passing (Carroll, 1945; Ferguson, 1941; Wherry and Gaylord, 1944). For purposes of item analysis or of analysis of the interrelations of measures in a single sample, the biserial or tetrachoric coefficients are more appropriate, since the difficulty of the items can be established by the proportion of correct responses.

When the qualitative predictor variable contains three or more categories, the multiserial *eta* coefficient, with the categories assigned scale values equal to their means on the continuous criterion variable, is recommended for purposes of prediction (Bittner, 1945; Wherry and Taylor, 1946). When these scale values are used, this coefficient is equal to the product-moment correlation between the two variables. One degree of freedom for each category is lost, provided the categories are ordered and weighted in terms of the differences in the criterion means of the category samples.

THE CLASSIFICATION PROBLEM

In the preceding sections selection has been treated as two interrelated activities: classification and prediction. The present section is concerned with the specific aspects of the classification problem as they arise in the development of the criterion and of the predictor "variables."

Detailed discussion of the development of classification procedures for educational, governmental, and military applications have been presented by Adkins (1947), Crawford and Burnham (1946), Davis (1947*a*), Flanagan (1948), Guilford and Lacey (1947), Melton (1949), Stead, Shartle, et al. (1940), Stuit (1947*a*), and Thorndike (1949). Earlier and somewhat more general discussions by Bingham (1937), Horst (1941), Hull (1928), Symonds (1931), and Viteles (1932) provide valuable reference materials. Additional publications of special interest, other than the numerous journal articles, include the critical reviews in the several *Mental Measurement Yearbooks* (Buros, 1949), the *Manual of Examination Methods* (Board of Examinations, 1937), and the *Assessment of Psychological Qualities by Verbal Methods* (Vernon, 1938).

Statement of objectives. A first step in selection is the determination of the characteristics to be used in classifying individuals. For each classification we must determine (1) what traits or performance characteristics are to be evaluated, (2) what standards of success are to be used, (3) what attributes should be present or absent, and (4) what relative importance is to be assigned to the characteristics. When these aims have been defined by the pooled judgments of competent persons, general agreement can be expected on the areas of performance to be included in the success continuum (Flanagan, 1948).

For simple activities the problem of selection is straightforward enough. But when success is defined as a "flying officer competent as a leader and as a pilot or navigator," we face a difficult problem of analysis. Success so defined, although considered as a unitary variate, is actually multidimensional (Bechtoldt, 1947; Horst, 1941; Toops, 1944). It is rare, therefore, that assessments on a single trait will provide the required ranking of the individual with respect to the total success variable.

Job analyses. In locating the areas of performance to be assessed, the selection psychologist uses several methods to secure the information he requires. These methods of analyzing an activity are discussed in detail by Horst (1941, Chapter III), Hull

(1928, Chapter IX), Stead, Shartle, et al. (1940, Chapter X), Thorndike (1949, Chapter II), and Viteles (1932, Chapter IX).

The major hazards of job analysis are (1) the possible omission of relevant aspects of the activity, (2) the introduction of extraneous factors through the biases of observers, (3) the inability of the observers to make accurate descriptive observations, (4) the differential effect of experience or training on the performances of successful and unsuccessful individuals, and (5) the relative contributions of previous training to the performances of novices and experienced personnel.

The methods of collecting the data required for job analyses may be grouped, for convenience of exposition, into (1) those based on the judgments of individuals other than the investigator, and (2) those involving direct observations by the investigator.

Sources of information concerning an activity include the published literature on the problem, records of performances of individuals, reports of causes of failure and of common complaints, formal job classification materials, and technical or training manuals. Other sources include interviews with experts, supervisors, and independent observers. Valuable data are often had from interviews with individuals engaging in the activity, including those just entering the field, those still in the learning phase, and those who have been judged relatively successful or unsuccessful. These sources must not be used uncritically.

Of more value are the direct observations by trained persons of individuals engaged in the activity and of situations characteristic of the activity. Systematic, recorded observations directed by a variety of hypotheses are usually necessary to determine those traits or characteristics in which the successful and unsuccessful individuals differ most significantly. Participation in the activity is often helpful to the investigator.

General forms for the summarizing of job duties, essential knowledges, specific information, characteristic activities, and typical inadequacies have been described (Adkins, 1947; Stead, Shartle, et al., 1940). Such forms or trait lists may prove to be deceptively exact, because operational definitions of traits such as initiative or emotional stability are by no means precise.

Tests of ability and skill. A test is defined as one or more tasks presented to the individual, together with the method of appraising the response (Thurstone, 1947). For each task some standard is provided by which to appraise the performance and in terms of which the individual is assigned to a specified class. In the testing procedures used in selection, research workers sometimes concentrate so exclusively on the "accuracy" of a response that they ignore other relevant properties. Such matters as variability, speed, frequency of omissions, frequency of wrong responses, and the ratio of correct to total responses may be indicative of behavior tendencies important in prediction (Coombs, 1948; Guilford and Lacey, 1947).

Tests are most frequently presented in printed form. They may utilize photographs, drawings, maps, etc., as well as verbal symbols, and may require that the subject provide definitions, state purposes, indicate causes or effects, recognize errors, evaluate alternatives, point out differences, rearrange materials according to some requirement, or indicate a common principle (Adkins, 1946; Mosier, Myers, and Price, 1945). Examples of the printed aptitude and achievement tests are provided by Brigham (1932), Crawford and Burnham (1946), Davis (1947*a*), Guilford and Lacey (1947), and by the publications of the several commercial testing organizations.

For large-scale testing programs, items calling for a limited response or a simple recognition are favored over items that require an essay type of answer. The limited response makes scoring objective; it minimizes the possible effects of fluency; and it reduces the ambiguity of the task set for the subject. The construction of such items is discussed by Adkins (1947) and by the staff of the Board of Examinations (1937). Considerable ingenuity, experience, and technical preparation, bolstered by a thoroughly systematic procedure is necessary for the development of items that require more than trite verbalizations and the parroting of definitions.

Apparatus tests of the work-sample type are utilized whenever the trait to be measured is defined in terms of manipulatory skill. Examples of apparatus tests are those that require the manipulation of peg or form boards, the solution of mazes and assembly tasks, or the judgment of speed and direction of movement. The tasks may require the use of slide or motion-picture projectors, reaction-time equipment, complex "training devices," or actual pieces of military or industrial equipment. Instructions on the development of apparatus tests are presented by Adkins (1947), Melton (1949), and Stuit (1947*a*).

The demands of expediency and convenience in large-scale testing are such that apparatus tests tend to be used only when forms that are easier to administer and to score are unacceptable. The problems of large-scale administration of apparatus tests were solved fairly satisfactorily, however, by the military services (Flanagan, 1948; Thorndike, 1947*a*; Stuit, 1947*a*). The military testers found it necessary to coordinate the testing sessions, to duplicate apparatus, to centralize controls and recording devices, and to standardize instructions.

Attitude and interest questionnaires. Considerable attention has been paid to attitude questionnaires and preference inventories because of the administrative convenience of these devices. The development of such questionnaires is discussed by Maller (1944), Symonds (1931), and Vernon (1938). Pertinent triennial reviews of the general topic of "Psychological Tests and Their Uses" are provided by the *Review of Educational Research.*

Ellis (1946) has pointed out that the validity studies of these questionnaires are equivocal. The authors of the devices usually find their instruments useful, but other investigators often fail to confirm their utility. Some success in military screening, where a psychiatric diagnosis was employed as the criterion, has been reported for several types of items (Ellis and Conrad, 1948; Guilford and Lacey, 1947; Wexler, 1947). However, the usefulness of questionnaires as measures of psychological variables has been questioned by Maller (1944) and by a sample of 79 psychologists polled by Kornhauser (1945), among others. The major objections to these devices appear to be the lack of internal consistency in the set of items, the influence of changes in mental set on the part of the subjects, the instability of the responses over samples of items and periods of time, and the absence of significant relations between such devices and other aspects of behavior.

Improvements designed to reduce the ease with which the "best" responses can be determined by the subject include the use of "forced-choice" items, with elements from two or more different continua presented in pair or triad form, and the application of empirical scoring weights based upon the performances of unlike groups (Fowler, 1947; Horst, 1941; Jurgensen, 1944; Meehl, 1945; Sisson, 1949; Wexler, 1947). Improvements in the stability of the validation data can be expected from the consistent use of cross-validation procedures. Modification of the usual criterion measures, so that they may be made to include evaluations of social and personal "adjustment" and of job satisfaction, may result in increased validity indices for these devices (Maller, 1944).

Biographical or personal history forms. Previous experiences of the subjects are often collected by self-rating or biographical data sheets as well as by the case history and interview methods to be considered later. Although administratively convenient, these self-rating procedures have been characterized by low validity in cross-validation. However, biographical items have proved moderately useful in the prediction of success in selling life insurance, in the selection of army officers and pilots, and in the admission of students to some educational institutions (Bittner, 1945; Guilford and Lacey, 1947; Kurtz, 1941; Richardson, 1947). The importance of the empirical determination of the weights to be assigned to the items and the establishment of their stability on new samples should be stressed, for there is a widespread uncritical use of biographical items scored on the basis of small-sample results or *a priori* judgments.

Personal history data can be criticized on two counts: (1) the ease with which the responses may be biased to the advantage of the respondent; and (2) the inability of the individual, even when favorably disposed, to recall accurately his past experiences. The procedure is most successful when the probability of falsification is low, when the responses are not regarded as accurate sources of personal history data, and when item-scoring weights are empirically determined.

Ratings. In selection, we usually want classifications based on the attributes of the subject and not on those of a rater observing the subject. In the rating processes, however, the data obtained are the responses of an observer to a situation of which the subject and his behavior are only a part. The observer, furthermore, is a complex mechanism subject to both systematic and variable errors. In spite of these weaknesses, ratings of one type or another are widely used in the development of criterion and predictor categories. Even such matters as salary and academic grades are basically ratings (Jenkins, 1946; Patterson, 1946).

Two different functions may be served by raters, namely, recording and evaluation. The recording function is of particular importance in tasks that leave no permanent record as, for example, in interpersonal situations or complex performance tasks. The function of evaluation is less straightforward. Here the rater serves as a computer presumed to possess the ability to synthesize nonlinear data and to determine what sorts of observations are to be included. These synthetic evaluations, characteristic of summary ratings and clinical judgments, are apt to be affected by the biases of the observers (Thorndike, 1949).

A variety of methods are used to secure an analytical judgment of some aspect of a performance or of a single trait, or to obtain a summary evaluation of the individual. These methods include paired comparisons, rank orders, and designations of position on a scale. Also, check lists on which specific aspects of the performance are marked as present or absent are used in rating a sequence of operations. The voting or nominating technique, as one form of rating device has been called, may be used when a number of judges are available for evaluating complex traits.

The problems involved in rating procedures are mainly those of securing discriminating, valid, and reliable measures. Discussions of the problems of developing such rating measures are contained in Symonds (1931) and Vernon (1938). For additional phases of the problems associated with ratings of limited behavior units and summary ratings used as criterion measures, see Adkins (1947), Cooper (1940), Flanagan (1948), Stuit (1947*a*) and Thorndike (1949).

The observer's biases pose the most difficult problem in the rating techniques. It is possible to reduce these biases by training the raters, by providing detailed descriptions of the performance to be evaluated, by specifying the standards to be used in the evaluation, and by indicating the ways in which the separate aspects are to be combined. Pooling the independent observations of several raters and securing repetitions of the evaluations are other means of counteracting bias.

Special mention should be made of the synthetic evaluations characteristic of clinical situations and interviews. The interview routinely used in selection is an extremely complex activity in which the interviewer can be considered as (1) a variable part of the social situation, (2) a recorder of the specific types of behavior being evaluated, and (3) an interpreter of the observations (Flanagan, 1948; Rundquist, 1947; Sarbin, 1944). In the interview, the problems of variability between individuals over samples of behavior and over a period of time may be entirely insignificant compared to the variability attributable to the situation and to the interviewer.

Adequate studies have not been made of the contribution of the evaluative type of interview to the efficiency of the selection process. Those that have been made indicate that the interview contributes relatively little to the other available techniques unless it is carefully standardized, uses trained interviewers, and is directed toward traits not otherwise evaluated (Davis, 1947*b*; Flanagan, 1948; Rundquist, 1947; Sarbin, 1944; Stuit, 1947*a*).

For vocational and military selection, there is little evidence of a significant increase in accuracy of prediction because of the addition of clinical judgments. Davis (1947*b*) summarized a number of studies using clinical tests and case history materials by saying that the subjective evaluation of empirical data appears to add little or nothing to the accuracy with which personnel can be classified for selection on the basis of suitable objective tests. The studies he reviewed include those made for the Coast Guard, the Army Air Forces, the Navy, and the Civil Aeronautics Authority. These studies utilized the Rorschach test, various work methods, and clinical evaluations of other types. A number of other investigators, working in situations in which the accuracy of their predictions could be tested, have likewise concluded that clinical judgments are of little use in classification techniques (Sarbin, 1944; Stuit, 1947*a*; Wallin, 1941).

It should not be inferred, however, that the interview is valueless in the selection process. The interview is a method of collecting biographical and preference data; it also provides an opportunity to evaluate the voice, the manner of expression, and the poise of an individual. Whether the interview is the most economical, efficient, and accurate method of securing such data depends upon the situation and upon the skill of the interviewer. It may also provide an opportunity for the interviewer to explain matters to those being selected, and to establish desirable personnel relations.

CRITERION MEASURES

An acceptable criterion of success is crucial because it constitutes the basis for validation. It provides the standard in terms of which the relevant predictor variables can be isolated, the efficient testing procedures separated from the inefficient, and the relative weights determined for use in predicting future performances and in combining sets of observations. The criterion must provide an adequate definition of the success continuum for the activity in question.

The success measure is usually assumed to lie on a single continuum. However, as Toops (1944) has demonstrated, success is not unitary; an analysis of any but the simplest activity will indicate that success is often the resultant of a large number of separate abilities, traits, skills, and knowledges. We can resolve this difficulty if we express the success continuum as a multidimensional variable, with each dimension an independent component. The success measure can then be defined by (1) a single overall evaluation, (2) a weighted composite of the separately measured components, and (3) a pattern or profile of these several variables (Adkins, 1937; Bechtoldt, 1947; Horst, 1941; Toops, 1944). The first two procedures are the ones most used.

The merits of criterion measures are judged in terms of their validity, reliability, and discrimination.

The validity of sets of criterion measures is evaluated in terms of statistical evidence of positive intercorrelations among the measures. The presence of significant correlations cannot safely be assumed; some of the psychologists working on military and industrial selection have found that individual criteria may be quite independent of one another, or, what is more pertinent, even independent of, if not negatively related to, the ultimate criterion (Flanagan, 1948; Jenkins, 1946; Stuit, 1947*a*).

The reliability and discrimination of criterion measures are evaluated in terms of statistical evidence of consistencies over periods of time and over samples of situations and by the distributions of the assigned values. Although no test can consistently predict a criterion that has zero reliability, high reliability in a criterion is desirable but not necessary. Low reliability introduces chance factors that attenuate the relations but do not introduce systematic irrelevant variables (Thorndike, 1949). Likewise, multiple classification categories of success permit improvements in prediction, but a dichotomous criterion, if relevant and reliable, will be superior, for validation purposes, to a multiple-category criterion that is either less valid or less reliable. An acceptable working procedure seems to be to utilize all the categories that (1) are necessary for the purpose of the selection situation, (2) can be shown to be reliably discriminated, and (3) meet the demands for accuracy of representation of the data.

THE PREDICTION PROBLEM

After the criterion and predictor classifications have been established, the next phase is the determination of the empirical relations between them and the verification on a new sample of the forecasts based upon these relations. These matters have been discussed by Adkins (1947), Flanagan (1948), Horst (1941), Hull (1928), Stead, Shartle, et al. (1940), Stuit (1947*a*), and Thorndike (1949). These general references form the basis for the following discussion of three major topics, namely: (1) the evaluation of prediction statements, (2) the factors influencing the accuracy of prediction, and (3) differential prediction.

Evaluation of Prediction Statements

Since predictions are based on probability estimates, errors of prediction are to be expected. For the *qualitative* case, the efficiency of prediction can be expressed in terms of the proportion of individuals correctly assigned (Guttman, 1941). The accuracy of *quantitative* predictions, on the other hand, is usually expressed as a function of the correlation between the criterion and the predictor measures.

The effectiveness of a given value of a correlation coefficient is usually expressed in terms of k, the coefficient of alienation, and E, the index of forecasting efficiency (Hull, 1928; Horst, 1941). Validity coefficients, corrected for attenuation, of less than 0.45 or 0.50 evaluated in terms of these indices are considered by some workers to be of little value in prediction unless a favorable selection ratio is obtained.

This conservative view is appropriate when a criterion value is to be predicted for each individual and the accuracy of the prediction for all cases is to be evaluated. However, for selection purposes, the utility of a prediction may be assessed in terms of the relative number of correct predictions and in terms of attaining or exceeding some critical value of the criterion (Horst, 1941). A higher validity coefficient means an increase in the proportion of cases scoring above some arbitrary criterion value. The magnitude of this increase for the special case of linear relations in a normal bivariate distribution can be estimated from the tables provided by Peters and Van Voorhis (1940). The correlation coefficient can be regarded as a direct index of the efficiency of prediction, according to Brogden (1946*a*). He shows that, when the regression is linear and the frequency distributions are similar, the correlation coefficient represents the ratio of the mean criterion score of the group selected from the top portion of the "combined predictor" distribution to the mean value that would be obtained by selecting a group of similar size by means of the *criterion* itself. Still another method of indicating the effectiveness of a selection device is provided by Richardson (1944), who defines predictive efficiency in terms of the increase in efficiency due to the use of a selection device, as compared with selecting the cases at random (Jarrett, 1948).

These measures refer to the accuracy of prediction in the original sample and not necessarily to the accuracy of the procedure when applied to a new sample. The value of a prediction procedure must be demonstrated on the new sample (Horst, 1941). The ratio of the quadratic mean of the errors of prediction in the new sample (based on the regression weights from the initial sample) to the standard error of estimate of the original sample provides an estimate of the accuracy in new samples. The quadratic mean of the errors is used, since both additive and multiplying constants based on the original sample are required. The correlation between predicted and obtained scores in the new sample is a less effective measure of accuracy because variations in the means and standard deviations are automatically corrected by the correlation coefficient (Horst, 1941).

The use of a cross-validation sample for the evaluation of the stability of a prediction equation is desirable because of the often unwarranted assumptions of (1) random sampling, (2) normality of the distribution, (3) independent errors, and (4) constant marginal frequency distributions on the predictor variables that are used to provide estimates of the sampling fluctuations of various regression statistics (Guttman, 1941). The mathematical solutions for nonnormal population distributions are not yet available.

Factors Influencing the Accuracy of Prediction

The accuracy of predictions is affected by a number of factors other than the intrinsic relation (validity) between the predictor variables and the criterion. In addition to the reliability of the classification procedures, such factors include (1) the selection ratio, (2) the difficulty level of the activity, (3) the method of selection, (4) the method of evaluation, (5) the representativeness of the sample, and (6) the number of measures used.

Selection ratio. No selection problem exists unless some individuals are to be chosen and some rejected. The ratio of the number chosen to the number available is the *selection ratio.* The effectiveness of a selection procedure in terms of the performance level of those accepted will be *inversely* related to the magnitude of this ratio. When the selection ratio is low (when only a few individuals are to be accepted), moderate to low validity coefficients may prove useful. On the other hand, if only a few individuals can be *rejected,* a much higher validity would be required for the same effectiveness.

Difficulty level of activity. The possible advantage of selection over nonselection varies *inversely* with the proportion of satisfactory individuals in a random sample. If, for example, 80 per cent of the candidates can be expected to succeed anyhow, there is little value in a selection program unless the proportion of successful individuals among those selected can be significantly increased over the 80-per-cent value (Toops, 1945*b*). When, however, the proportion of successful candidates in a nonselected group is 50 per cent or less, measures having validity coefficients as low as 0.30 may prove valuable.

The theoretical effectiveness of a selection procedure as determined by difficulty level, selection ratio, and obtained validity has been investigated by Taylor and Russell (1939) both in terms of the expected proportions of successful candidates among those selected and under the assumption of linear relations and normally distributed variables. Their results indicate that acceptable efficiency can be achieved even though the index of forecasting efficiency is below the 10-per-cent point recommended by Hull (1928) as a critical level.

Method of selection. Selection may involve a single hurdle or successive hurdles. In the single-hurdle method, all individuals are evaluated on all the selection devices; with successive hurdles, the number of applicants is reduced successively by separate operations until at the final hurdle only a few individuals need be considered. Successive hurdles are often found in educational programs where there are periodic rejections for unsatisfactory performance. This procedure can be criticized, however, when there is a drastic use of cutting scores on separate tasks. It is sometimes assumed that failure to perform well on each task warrants elimination, and that high performance on one measure cannot compensate for failure on another. As Toops (1932) has pointed out, this disadvantage can be overcome to some extent by a proper arrangement of the tasks and the use of relatively high cutting scores in the first measures. The tasks should be arranged in decreasing order of validity, the most valid first, etc. At the final stage, the selection problem may be of little consequence, since the group may by then be quite homogeneous.

Method of evaluation. The predicted criterion scores may be obtained in terms of formal measures (linear regression equations, cutoff scores, matched profiles) or in terms of a clinical evaluation of the available data.

The more efficient procedures utilize all available data, and the predictions are made by a formal method. The linear regression equation and the multiple correlation technique are usually applicable (Thorndike, 1949). Since this process is basically one of determining the "best" weights in a linear equation, the special problems of this process will be considered later in the section on the combining of measures.

The methods of successive and multiple cutting scores and of matched profiles have been recommended for use whenever (1) the relations between the criterion and predictor measures are conspicuously nonlinear, (2) competence in one area cannot compensate for weakness in another, or (3) a specified pattern of desirable and undesirable traits has been established (Ruch, 1945; Toops, 1945*b*). A convenient method of utilizing these nonlinear relations or critical values, in case they exist, is to establish a cutting score on each of the crucial attributes. Individuals who fall below the success score are then automatically eliminated, regardless of their standing on other variables. The multiple-cutting-score method selects the most efficient pattern of cutoff values from a series of possible combinations and eliminates individuals falling outside this pattern.

These cutoff methods and the linear regression techniques may not result in the selection of the same individuals (Thorndike, 1949). Individuals barely above the cutting score on each of two predictor variables may be rejected by the linear regression or "summation of traits" method. Those below the cutting score on one of the variables, but who score high on the other measure, may be accepted by some linear composite selection method, but rejected by the successive cutoff procedure.

Additional work needs to be done on the relative accuracy of the predictions made by the methods described above. For large-scale selection problems in which approximately linear relations can be found, the composite score method is probably most appropriate. Critical cutting scores on the composite criterion are of greatest value when the proportion of individuals to be selected and the proportion above a given point on the composite score distribution can be determined.

Representativeness of samples studied. Perhaps the most insidious factor in selection is the nonrepresentativeness of the validity relations determined on a sample of individuals (Burt, 1944; Marks, 1947; Flanagan, 1948; Thorndike, 1949). Systematic biases may be introduced by the sampling procedure. The ideal sampling procedures may be difficult to formulate, however, because of the uncertainty in the definition of the population from which the selection is to be made. The population to be sampled includes not only those individuals available at the moment but also those who will engage in the activity in the future. When the population cannot be defined accurately, an estimate of the sampling variability may be secured, with some degree of accuracy, from the empirical results of successive large samples.

Since the relations determined on a sample are to be applied to some type of population in a selection process, the heterogeneity of the sample as compared with that of the population is important, especially when successive hurdles are used. As long as the obtained coefficients are to be used with the sample or with similar samples, no correction for heterogeneity is necessary. However, if the selection procedure is to be applied to a more heterogeneous population, an estimate is required of the correlation that would have been obtained had the total population been permitted to reach the final stage of performance. The general formulation of this problem has been presented by Burt (1943), by Kelley (1947), and by Thorndike (1949).

Another sampling question arises from the use of special groups for validation purposes and from the effects of experience on performance. This type of sampling may introduce irrelevant factors into the prediction equation. Since training may modify differentially many of the traits used for prediction, it cannot be assumed that performances before and after a given amount of experience are comparable. A similar questionable assumption is frequently made in the validation of personality tests and questionnaires through the use of hospitalized and nonhospitalized groups (Wexler, 1947). These assumptions should not be made unless there exists positive evidence of their validity.

Number of predictor variables. The stability of predictions for new samples tends to decrease with an increase in the number of predictor variables in the original sample when the regression weights are determined by least squares. The addition of variables to the original set of predictors will not, however, reduce the accuracy of prediction in the original sample (Horst, 1941; Reed, 1941). In an effort to increase the stability of their predictions, psychologists try to reduce the number of predictor variables to a minimum. Two different procedures have been used for this purpose. For the prediction of a single complex criterion, batteries of two to five complex predictor variables of the broad "aptitude" type are developed. For the prediction of several criteria, different aptitude batteries would be required. Another solution attempts to develop relatively "pure" measures of the separate criterion "traits" and to combine in various ways the several tests from the single pool of predictor measures. This second procedure is defended on the basis of the increased probability of differential prediction (see the next section) as well as on the basis of general efficiency in test construction.

The importance of comprehensive coverage of all the relevant criterion traits, however, runs counter to the desirability of reducing the number of variables. The effectiveness of a selection procedure will be lessened if any significant aspect of the criterion is omitted from the predictor variables. This has been one of the frequent sources of inaccuracies of prediction (Horst, 1941).

Differential Prediction

Selection can be extended to the situation in which each individual is considered for assignment to one of several activities. This more general selection process has been variously termed *multiple selection, classification and differential prediction,* or *placement.* The present discussion will review the major points regarding this problem that have been raised by Brogden (1946*c*), Burt (1943), Flanagan (1947*a*, 1948), Horst (1941), and Thorndike (1949); namely, (1) the characteristics of the predictor variables used in differential prediction, and (2) the procedures used in effecting such classifications.

The basic assumption here is that a restricted set of trait measurements, differentially weighted, can be used to predict success in two or more areas of activity. For reasons of parsimony, the number of these fundamental measures should be as small as possible, and each one should be significantly related to only a few criterion measures (Horst, 1941).

From this general point of view, the tests developed for the prediction of single and multiple criteria can be contrasted. For the prediction of single criterion measures, complex tests of the miniature job-sample type resembling the criterion in content, materials, type, and complexity of task have proved efficient. These are designed to measure the hypothetical traits of the criterion in combinations resembling those found in the activity itself (Thorndike, 1949). The tests recommended for use in differential prediction, on the other hand, tend to be more nearly homogeneous measures of a single functional unity and have the advantage that, in combination, the proper weight of each trait for the prediction of each criterion can be determined. Such measures usually represent performances on thoroughly learned materials.

At the present time, two methods are used to determine the functional unities to be represented by these homogeneous tests. One involves the appraisal of the test materials by sophisticated individuals. The other method utilizes the intercorrelations among the measures and defines the functional unities in terms of high intercorrelations among those measures that may be combined, and low correlations between sets of such combined measures.

After the measures have been assembled and the regression weights for the prediction of each of the several criteria determined, the problem of the most efficient assignment of the individuals arises. The objective is to maximize the selection efficiency for the available group in terms of performance in the several activities (Brogden, 1946*c*; Thorndike, 1949). The attainment of this objective involves consideration of (1) the reliability of the differentiation between predicted scores as a function of the reliabilities of the measures and their intercorrelations, (2) the selection ratio and critical rejection scores, and (3) the relative importance of the activities.

For an accurate differentiation between activities to which individuals are assigned, the reliabilities of the separate scores should be comparable and as high as possible, and the intercorrelations of the measures should be minimized. If significant differences are to be located, the chance variations in individual scores should be small in relation to the between-score differences for the average individual. If the obtained differences are relatively small, as would be the case with correlated composite scores or with uniformly low multiple correlations, then relatively high reliability is required (Thorndike, 1949). The ratio of the standard deviation of a difference (in standard score form) for a given *individual* to the standard deviation of standard score differences for the *group* has been suggested for the evaluation of pairs of profile differences used in differential prediction (Kelley, 1947; Bennett and Doppelt, 1948).

Although the validity of each composite score for its associated criterion should be as high as possible, only one of the several validity coefficients (for the different composite scores) for a single activity should be high, and all the other coefficients should be as low as possible (Thorndike, 1949). The highest coefficient should be that between the criterion measure of an activity and the composite score used to select individuals for that activity.

When the selection ratio is favorable and critical rejection scores can be established, a successive-approximation procedure for maximizing the overall effectiveness of selection is applicable. Brogden (1946*c*) has demonstrated that, in the linear case, efficient differentiation can be effected by establishing, first, a set of critical rejection composite scores for each activity in terms of the number to be accepted, and, second, a set of critical assignment scores equal, in each case, to the differences between the critical rejection scores. For the individuals above the critical values in two or more activities, the assignment is made to the activity for which the difference in predicted standard score units is greatest in terms of the critical difference scores. As the number of activities is increased, the complexity of this successive-approximation solution is increased.

A solution has been proposed for several special cases in which all individuals are to be assigned and no screening can be accomplished (Flanagan, 1948). Whenever individuals are to be assigned in equal numbers to equally important and equally difficult activities involving independent sets of predictor traits, each individual is assigned to the position for which his composite aptitude score is the highest. If cutting scores can be used, say, at the mean of the predicted scores, then for the case of three independent activities, seven-eighths of the individuals could be assigned to one activity with the expectation that their performance would be average or better. If some selection can take place, the individuals above the rejection values on two or more composite scores are assigned to the position for which their predicted scores are highest.

In the practical situation involving differences in the difficulty levels and in the relative importance of the activities, as well as in the number of individuals required, successive selection in terms of the importance of the activity may be sufficiently accurate (Flanagan, 1948). This general problem, however, has not yet been solved.

THE COMBINING OF MEASURES

In the selection process it is often necessary to combine several measures for the purpose either of defining a single composite criterion or of obtaining predictions of future performance. The solutions to this problem can be classified as: (1) the weighting of predictor (independent) variables given some single criterion (dependent) variate, which is the ordinary multiple-correlation procedure; (2) the weighting of measures when there is no dependent variable and the multiple-correlation procedures are inappropriate; (3) the simultaneous weighting of sets of predictor measures and sets of criterion measures to effect the maximum correlation between two composite variables. The first two of these solutions are those most frequently used in selection problems. The third procedure has been criticized as inappropriate for practical selection problems (Horst, 1941; Thorndike, 1949).

Weighting with the use of a dependent variable. Two situations that can be considered as a single problem are: (1) the prediction of a single criterion measure from a set of predictor variables, and (2) the combination of sets of partial criterion measures when an ultimate criterion is available. The general solution is to weight the several measures in such a way that the weighted linear composite will conform as closely as possible to the values of the single criterion observations. If the single criterion is a quantitative one, the theoretically best solution under the least-squares principle leads to a multiple-regression equation. If the criterion is a qualitative variate, the "best" combination in the same sense is provided by Fisher's discriminant function which, for a dichotomous criterion scored 0 and 1, can be regarded as a regression-equation problem. These techniques take into consideration the intercorrelations of the measures as well as their correlations with the criterion.

The computational difficulties of multiple correlation tend to restrict the applicability of this method. Dwyer (1945), Guttman (1941), and Hoel (1947), however, provide both theoretical and computational simplifications that should extend the usefulness of the technique. Rapid approximations to the exact solutions are provided by a modification of the Kelley-Salisbury iterative solution of the regression weights developed for use in the Army Air Forces research program (Thorndike, 1949). In practice, the exact weights determined by these methods may be modified for computational convenience.

Sets of more or less arbitrary, or intuitive, weights representing the judgments of experts have been used in combining predictor or criterion measures. In spite of the "subjectivity" of such weights, the stability of these solutions, as shown by the effectiveness in new samples, may be comparable to that of the more rigorous techniques (Reed, 1941). In any case the effectiveness in prediction rather than the method of obtaining the weights is the crucial point at issue.

Another important problem associated with the multiple-correlation technique is the "regression" effect in new samples as the ratio of the number of predictor variables to the number of cases in the samples increases (see above). The generality of this fact, together with the absence of a suitable criterion for the number of variables to be used, has led to the "rule of thumb" that multiple-correlation problems should be restricted to fewer than six predictor variables and that the regression coefficients, in any case, should be based upon large samples of individuals (Horst, 1941; Adkins, 1947; Thorndike, 1949). The actual determination of the stability of regression weights on new samples will provide the evidence for the justification of a larger number of variables.

The multiple-correlation techniques, or approximations thereto, also provide a basis for deciding which are the more "efficient" items of a test (Richardson, 1936*b*; Adkins and Toops, 1937; Flanagan, 1939*b*; Horst, 1941). The dependent variable may be either the total test score or an "external criterion" measure. For computational convenience, an assumption of equal or zero interitem correlations is often introduced, although the adequacy of the solution is thereby reduced. Detailed discussions of these techniques are given by Adkins (1947), Conrad (1944), Davis (1946*a*), Horst (1934*b*), and Toops (1941). The main objectives of item analysis are (1) to increase the internal consistency, as measured by the average interitem correlation, of the set of items, (2) to increase the predictive efficiency of some external criterion by a set of items, and (3) to locate specific faults in the construction of the test items.

Whenever the purpose is to secure an internally consistent set of items, as in the development of a measure of a single trait, the dimensionality of the set of item intercorrelations and the "homogeneity" of the responses should be considered. For the practical purposes of selection, useful approximations to this objective are obtained from the correlation of the item with the total score, since this coefficient is proportional to the average correlation of the item with all the items in the set (Richardson, 1941*a*). Rejection of items with low item-test coefficients will increase the average intercorrelation and, in this sense, the homogeneity of the set.

If the original set of items measures two or more factors, however, the final set of items selected on the basis of high item-test correlations may contain one, two, or more factors, depending on the number of items representing each factor in the original collection. There is no assurance that the final set of items will measure only a single trait (Mosier, 1936; Sletto, 1937). If the item selection is to be accomplished through the use of the item-test coefficient, an efficient solution would be realized if the homogeneous subsets of items representing the different factors in the test were first determined, and then the several scores on the subsets and the item-subtest correlation were used for the selection of the items (Wherry and Gaylord, 1943; Davis, 1947*a*).

For purposes of prediction, the relation between the items used as predictors and an external criterion is determined. This relation is usually represented either by one of the correlation coefficients or an approximation thereto, or by some type of regression coefficient (Davis, 1946*a*; Adkins, 1947; Thorndike, 1949). One procedure treats each item as an individual test and neglects the interitem correlations. More efficient procedures using all the relevant data have been developed by Horst (1934*b*) and Toops (1941).

The sampling variation in item-discrimination statistics is appreciable even for samples as large as 400 cases, although the difficulty indices of items may be fairly stable (Davis, 1946*b*; Travers, 1942). Furthermore, unless the distribution of validity coefficients is significantly greater than that expected by chance, there is little reason to select any particular sample of items from the total test (Merrill, 1937). These sampling problems are sufficiently serious to warrant a questioning of the often mechanical application of item-analysis techniques in test construction.

Weighting without the use of a dependent variable. The basic discussions of this linear weighting problem are presented by Burt (1936), Kelley (1947), Richardson (1941*a*), and Wilks (1938). In general, there is no one "best" method of weighting; the different methods accomplish different objectives. The methods may involve the simple addition of arbitrary numerical values, the use of values representing the judgments of experts, or some function of the intercorrelations of the scores and of their reliability coefficients.

The distinction between effective and nominal weights is important to the weighting problem. The nominal weights are the coefficients, W_j, of the variables in the linear equation, $T_i = \sum_j W_j X_{ji}$. The effective (or functional) weights have been defined as "the proportion of the total variance of the composite T_i that is contributed by the particular (weighted) variable $W_j X_{ji}$" (Richardson, 1941*a*). The effective weight of a variable depends upon the nominal weight of that variable and its standard deviation as well as upon the weighted correlations of the variable with each measure (including itself) in turn. The correlations are weighted by the dispersions of the several measures, and the self-correlation is defined as unity.

The simple addition of scores, as in the case of a set of test items, is sufficiently accurate for the combining of large numbers of variates. The rationale for this simple procedure is that, as the number of positively correlated variables increases, the correlation between any two sets of weighted scores approaches unity and the effect of differential weighting tends to disappear. However, if the number of measures to be added together is not large, the dispersions and intercorrelations of the measures will influence significantly the effective weighting.

The method of weighting scores inversely as their standard deviations is a common practice. If the tests are considered to be of equal importance and the dispersions to be artifacts of the arbitrary characteristics of the test, then the addition of the *standard scores* of the individuals is appropriate. Richardson (1941*a*) has pointed out two important characteristics of this method. In the first place, if the number of variables is three, or greater, the variance contributions of the several variables to the composite are not necessarily equal, even though standard scores are added (i.e. the nominal weights are unity). Differential weighting will occur unless the variables are equally correlated. The second point is that, in so far as the differences in the variances of the measures reflect differences in internal consistency (and reliability, in this sense), the addition of standard scores will, in effect, increase the relative weight of the less reliable tests, since the variance of the composite scores is directly related to the internal consistency of the set of measures.

If the set of measures can be regarded as parallel forms of a single test, such that the intertest correlations corrected for attenuation are unity, the measures may be weighted in terms of a function of their respective reliabilities (Kelley, 1947). This procedure is justified on the grounds that an increase in the size of the sample increases the reliability of the composite. The effective contribution of each measure so weighted to the composite is directly proportional to its reliability coefficient and inversely proportional to its error variance (Richardson, 1941*a*).

A related problem is that of weighting the measures so that the reliability of the *composite* will be maximal. The solution, which deals with a property of the composite, is not the same, however, as that obtained when the assumption is made that the intrinsic correlation coefficients are unity. Thomson (1940) has shown that weighting for the case of maximum battery reliability is a function of the reliabilities and of the intercorrelations of the measures.

Another theoretically important method of determining a set of weights is applicable to the combination of positively correlated attributes that are considered as separate aspects of a single variable. The same solution can be arrived at through three approaches; the scores may be weighted so that: (1) the dispersion of the composite is as large as possible (Horst, 1936); (2) the variations between the weighted scores for each individual are minimized (Edgerton and Kolbe, 1936), and (3) the composite score for each individual best represents his standing on all the measures of the composite (Burt, 1936; Hotelling, 1933). Horst (1941) indicates that the units of measurement must be considered in these (principal axes) solutions. Since the least-squares determination of the principal axes requires the solution of the characteristic equation, the problem becomes computationally cumbersome for more than five or six measures. However, this weighting procedures has a sound rationale and is considered appropriate for the combination of a series of measures that represent different aspects of the underlying variable.

The computational difficulties associated with the principal-axes solution have led to the introduction of weighting in terms of the correlation of the variable with the average centroid axis of the system or according to the contribution to the total variance of the composite (Burt, 1936; Richardson, 1941*a*). Burt (1936) points out that another approximation can be secured by assuming that only one variable is being measured and by using a single-factor solution for the weights. These solutions will not differ greatly, but the accuracy of the centroid weighting as an approximation to the method of maximum variance may not be close (Edgerton and Kolbe, 1936). The centroid method is computationally feasible for relatively large numbers of variables for which either the intercorrelations of the measures or their correlations with the sum of the variables are available. The item-analysis procedures, using the item-test correlation coefficient (an internal criterion), represent applications of this method (Richardson, 1941*a*).

Another method of weighting measures representing various aspects of success utilizes the judgment of a "competent" group of individuals regarding the relative importance of the several elements (Horst, 1941;

Toops, 1945*b*). These judgments can be expressed in a number of ways. If there are two or more judges, each may be required to apportion among the several components of the activity a given set of numbers or "bids." The number assigned to each element represents the judge's evaluation of the importance of that aspect. The judges, in turn, can be weighted by some estimate of their competence or experience. The final set of weights can then be reduced proportionately to any desired total. The bids, in turn, may be adjusted for the intercorrelations and reliabilities of the measures. These "rational" weights can now be considered as representing the desired effective contributions of the separate measures to the composite value, but the nominal weights must still be determined. The determination of the nominal weights requires the analysis of a set of simultaneous equations that express the nominal weights as a function of the dispersions and intercorrelations of the variables and of the effective or rational weights (Horst, 1941; Bechtoldt, 1947). The error often made in the use of rational or intuitional weights lies in considering these judgments as nominal weights rather than as the effective weights.

REFERENCES*

The following references were selected on the basis of accessibility to students, recency of publication, and completeness of treatment; priority of publication was not considered.

Adkins, D. C. Test construction in public personnel administration. *Educ. psychol. Measmt.*, 1944, **4**, 141–160.

Adkins, D. C. Construction and analysis of written tests for predicting job performance. *Educ. psychol. Measmt.*, 1946, **6**, 195–211.

Adkins, D. C. *Construction and analysis of achievement tests.* Washington, D. C.: Superintendent of Documents, 1947.

*Reports prepared under contract with the Office of Scientific Research and Development usually bear an OSRD number. In addition, many carry a PB number, which indicates that they may be obtained from the Office of Technical Services, U. S. Department of Commerce, Washington 25, D. C.

Adkins, D. C. Needed research on examining devices. *Amer. Psychologist,* 1948, **3**, 104–106.

Adkins, D. C., and H. A. Toops. Simplified formulas for item selection and construction. *Psychometrika,* 1937, **2**, 165–171.

Alexander, H. W. The estimation of reliability when several trials are available. *Psychometrika,* 1947, **12**, 79–99.

Anastasi, A. The nature of psychological 'traits.' *Psychol. Rev.*, 1948, **55**, 127–138.

Babitz, M., and N. Keys. A method for approximating the average intercorrelation coefficient by correlating the parts with the sum of the parts. *Psychometrika,* 1940, **5**, 283–288.

Baier, D. E. Selection and evaluation of West Point cadets. *Educ. psychol. Measmt.*, 1948, **8**, 193–199.

Bechtoldt, H. P. Problems in establishing criterion measures. In D. B. Stuit (Ed.), *Personnel research and test development in the Bureau of Naval Personnel.* Princeton: Princeton University Press, 1947.

Bechtoldt, H. P., J. W. Maucker, and D. B. Stuit. The use of order of merit rankings. In *New methods in applied psychology.* College Park: University of Maryland, 1947. Pp. 26–33.

Bellows, R. M. Procedures for evaluating vocational criteria. *J. appl. Psychol.*, 1941, **25**, 499–513.

Bennett, G. K., and J. E. Doppelt. The evaluation of pairs of tests for guidance use. *Educ. psychol. Measmt.*, 1948, **8**, 319–325.

Bergmann, G., and K. W. Spence. The logic of psychophysical measurement. *Psychol. Rev.*, 1944, **51**, 1–24.

Bingham, W. V. *Aptitudes and aptitude testing.* New York: Harper, 1937.

Bingham, W. V., and B. V. Moore. *How to interview.* (Revised Ed.) New York: Harper, 1941.

Bittner, R. H. Quantitative prediction from qualitative data: Predicting college entrance from biographical information. *J. Psychol.*, 1945, **19**, 97–108.

Board of Examinations. *Manual of examination methods.* Chicago: University of Chicago, 1937.

Bordin, E. S. A theory of vocational interests as dynamic phenomena. *Educ. psychol. Measmt.*, 1943, **3**, 49–65.

Bridgman, P. W. *The logic of modern physics.* New York: Macmillan, 1932.

Brigham, C. C. *A study of error.* Princeton: College Entrance Examination Board, 1932.

Brogden, H. E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. *J. educ. Psychol.*, 1946*a*, **37**, 65–76.

Brogden, H. E. Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. *Psychometrika,* 1946*b*, **11**, 197–214.

Brogden, H. E. An approach to the problem of differential prediction. *Psychometrika,* 1946*c*, **11**, 139–154.

Buros, O. K. (Ed.). *The third mental measurements yearbook.* New Brunswick: Rutgers University Press, 1949.

Burt, C. The analysis of examination marks, memorandum I. In P. Hartog, E. C. Rhodes, and C. Burt, *The marks of examiners.* London: Macmillan, 1936. Pp. 245–314.

Burt, C. Validating tests for personnel selection. *Brit. J. Psychol.,* 1943, **34,** 1–19.

Burt, C. Statistical problems in the evaluation of army tests. *Psychometrika,* 1944, **9,** 219–235.

Burt, C. The reliability of teachers' assessment of their pupils. *Brit. J. educ. Psychol.,* 1945, **15,** 80–92.

Burt, C. Symposium on the selection of pupils for different types of secondary schools. I. A general survey. *Brit. J. educ. Psychol.,* 1947, **17,** 57–71.

Carroll, J. B. The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika,* 1945, **10,** 1–19.

Carter, L. F., and F. J. Dudek. The use of psychological techniques in measuring and critically analyzing navigators' flight performance. *Psychometrika,* 1947, **12,** 31–42.

Conrad, H. S. Characteristics and uses of item-analysis data. OSRD Report 4034, Aug. 1944 (PBL 13296).

Coombs, C. H. Some hypotheses for the analysis of qualitative variables. *Psychol. Rev.,* 1948, **55,** 167–174.

Cooper, J. H. Rating forms. In W. H. Stead, C. L. Shartle, et al., *Occupational counseling techniques.* New York: American, 1940.

Crawford, A. B., and P. S. Burnham. *Forecasting college achievement.* New Haven: Yale University Press, 1946.

Cronbach, L. J. Response sets and test validity. *Educ. psychol. Measmt.,* 1946, **6,** 475–494.

Cronbach, L. J. Test reliability: Its meaning and determination. *Psychometrika,* 1947, **12,** 1–16.

D'Abro, A. *The decline of mechanism (in modern physics).* New York: Van Nostrand, 1939.

Darley, J. G. A study of clinical predictions of student success or failure in professional training. *J. educ. Psychol.,* 1938, **29,** 335–354.

Davis, F. B. *Item analysis data: Their computation, interpretation, and use in test construction.* Cambridge: Harvard Graduate School of Education, 1946*a*.

Davis, F. B. Notes on test construction: The reliability of item analysis data. *J. educ. Psychol.,* 1946*b*, **37,** 385–390.

Davis, F. B. *The AAF qualifying examination.* Army Air Forces Aviation Psychology Program Research Report 6. Washington, D. C.: U. S. Government Printing Office, 1947*a*.

Davis, F. B. *Utilizing human talent.* Washington, D. C.: Commission on Implications of Armed Services Educational Programs, American Council on Education, 1947*b*.

Deemer, W. L. *Records, analyses and test procedures.* Army Air Forces Aviation Psychology Program Research Report 18. Washington, D. C.: U. S. Government Printing Office, 1947.

DuBois, P. H. *The classification program.* Army Air Forces Aviation Psychology Program Research Report 2. Washington, D. C.: U. S. Government Printing Office, 1947.

Dvorak, B. J. The new USES general aptitude test battery. *J. appl. Psychol.,* 1947, **31,** 372–376.

Dwyer, P. S. Recent developments in correlational technique. *J. Amer. statist. Ass.,* 1942, **37,** 441–460.

Dwyer, P. S. The square root method and its use in correlation and regression. *J. Amer. statist. Ass.,* 1945, **40,** 493–503.

Dyer, H. S. The differential prediction of college achievement. In *Exploring individual differences.* Washington, D. C.: American Council on Education, Series 32, 1948, **12,** 80–87.

Edgerton, H., and L. E. Kolbe. The method of minimum variation for the combination of criteria. *Psychometrika,* 1936, **1,** 183–187.

Ellis, A. The validity of personality questionnaires. *Psychol. Bull.,* 1946, **43,** 385–440.

Ellis, A. A comparison of the use of direct and indirect phrasing in personality questionnaires. *Psychol. Monogr.,* 1947, **61,** No. 284.

Ellis, A., and H. S. Conrad. The validity of personality inventories in military practice. *Psychol. Bull.,* 1948, **45,** 385–426.

Eysenck, H. J. Student selection by means of psychological tests—a critical survey. *Brit. J. educ. Psychol.,* 1947, **47,** Pt. I, 20–39.

Fearing, F., and F. M. Fearing. Factors in the appraisal interview considered with particular reference to the selection of public personnel. *J. Psychol.,* 1942, **14,** 131–153.

Ferguson, G. A. The factorial interpretation of test difficulty. *Psychometrika,* 1941, **6,** 323–329.

Ferguson, G. A. On the theory of test discrimination. *Psychometrika,* 1949, **14,** 61–68.

Festinger, L. The treatment of qualitative data by "scale analysis." *Psychol. Bull.,* 1947, **44,** 149–161.

Flanagan, J. C. *The Cooperative achievement tests: A bulletin reporting the basic principles and procedures used in the development of their system of scaled scores.* New York: Cooperative Test Service of the American Council on Education, 1939*a*.

Flanagan, J. C. General considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution. *J. educ. Psychol.,* 1939*b*, **30,** 674–680.

Flanagan, J. C. Current trends in psychology. Paper presented at the Conference on Current Trends in Psychology. University of Pittsburgh, Mar. 6, 1947*a*.

Flanagan, J. C. Units and norms in educational measurement. In *National project in educational measurement.* American Council on Education, Series 28, 1947*b*, **11,** 8–12.

Flanagan, J. C. *The aviation psychology program in the Army Air Forces.* Army Air Forces Aviation Psychology Program Research Report 1. Washington, D. C.: U. S. Government Printing Office, 1948.

Fowler, H. M. The consistency of items of an activity preference blank. *Psychometrika,* 1947, **12,** 221–232.

Freeman, G. L. Using the interview to test stability and poise. *Publ. Person. Rev.,* 1944, **5,** 89–94.

Garrett, H. E. The discriminant function and its use in psychology. *Psychometrika,* 1943, **8,** 65–79.

Guilford, J. P. New standards for test evaluation. *Educ. psychol. Measmt.,* 1946, **6,** 427–438.

Guilford, J. P. The discovery of aptitude and achievement variables. *Science,* 1947, **106,** 279–282.

Guilford, J. P., and J. I. Lacey. *Printed classification tests.* Army Air Forces Aviation Psychology Program Research Report 5. Washington, D. C.: U. S. Government Printing Office, 1947.

Guilford, J. P., C. Lovell, and R. M. Williams. Completely weighted versus unweighted scoring in an achievement examination. *Educ. psychol. Measmt.,* 1942, **2,** 15–21.

Gulliksen, H. The content reliability of a test. *Psychometrika,* 1936, **1,** 189–194.

Gulliksen, H. The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika,* 1945, **10,** 79–91.

Gulliksen, H. Paired comparisons and the logic of measurement. *Psychol. Rev.,* 1946, **53,** 199–213.

Guthrie, E. R. Psychological facts and psychological theory. *Psychol. Bull.,* 1946, **43,** 1–20.

Guttman, L. Mathematical and tabulation techniques. Supplementary study B. In P. Horst, *The prediction of personal adjustment.* New York: Social Science Research Council Bulletin 48, 1941.

Guttman, L. A basis for scaling qualitative data. *Amer. sociol. Rev.,* 1944, **9,** 139–150.

Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika,* 1945, **10,** 255–282.

Guttman, L. The test-retest reliability of qualitative data. *Psychometrika,* 1946, **9,** 81–95.

Guttman, L. Scale and intensity analysis for attitude, opinion, and achievement. In *New methods in applied psychology.* College Park: University of Maryland, 1947. Pp. 173–180.

Hoel, P. G. *Introduction to mathematical statistics.* New York: Wiley, 1947.

Horst, P. Increasing the efficiency of selection tests. *Person. J.,* 1934*a,* **12,** 254–259.

Horst, P. Item analysis by the method of successive residuals. *J. exp. Educ.,* 1934*b,* **2,** 229–244.

Horst, P. Obtaining a composite measure from a number of different measures of the same thing. *Psychometrika,* 1936, **1,** 53–60.

Horst, P. *The prediction of personal adjustment.* New York: Social Science Research Council Bulletin 48, 1941.

Horst, P. A generalized expression for the reliability of measures. *Psychometrika,* 1949, **14,** 21–32.

Hotelling, H. Analysis of a complex of statistical variates into principal components. *J. educ. Psychol.,* 1933, **24,** 417–441, 498–520.

Hotelling, H. The most predictable criterion. *J. educ. Psychol.,* 1935, **26,** 139–142.

Hotelling, H., P. A. Sorokin, L. Guttman, and E. W. Burgess. The prediction of personal adjustment: A symposium. *Amer. J. Sociol.,* 1942, **48,** 61–86.

Hoyt, C. Test reliability obtained by analysis of variance. *Psychometrika,* 1941, **6,** 153–160.

Hull, C. L. *Aptitude testing.* New York: World, 1928.

Hull, C. L. *Principles of behavior.* New York: Appleton-Century-Crofts, 1943.

Jackson, R. W. B., and G. A. Ferguson. *Studies on the reliability of tests.* Toronto: Department of Educational Research, University of Toronto, 1941.

Jackson, R. W. B., and G. A. Ferguson. A plea for a functional approach to test construction. *Educ. psychol. Measmt.,* 1943, **3,** 23–28.

Jarrett, R. F. Per cent increase in output of selected personnel as an index of test efficiency. *J. appl. Psychol.,* 1948, **32,** 135–145.

Jeffreys, H. *Scientific inference.* Cambridge: Cambridge University Press, 1937.

Jenkins, J. G. *Psychology in business and industry.* New York: Wiley, 1935.

Jenkins, J. G. Validity for what. *J. consult. Psychol.,* 1946, **10,** 93–98.

Jurgensen, C. E. Report on the 'classification inventory,' a personality test for industrial use. *J. appl. Psychol.,* 1944, **28,** 445–460.

Kelley, T. L. *Fundamentals of statistics.* Cambridge: Harvard University Press, 1947.

Klein, G. S. Self-appraisal of test performance as a vocational selection device. *Educ. psychol. Measmt.,* 1948, **8,** 69–84.

Kogan, L. S. Analysis of variance—repeated measurements. *Psychol. Bull.,* 1948, **45,** 131–143.

Kornhauser, A. Replies of psychologists to a short questionnaire on mental test developments, personality inventories, and the Rorschach test. *Educ. psychol. Measmt.,* 1945, **5,** 3–15.

Kuder, G. F. The stability of preference items. *J. soc. Psychol.,* 1939, **19,** 41–50.

Kuder, G. F. Note on 'classification of items in interest inventories.' *Occupations,* 1944, **22,** 484–487.

Kuder, G. F., and M. W. Richardson. The theory of the estimation of test reliability. *Psychometrika,* 1937, **2,** 151–160.

Kurtz, A. K. Recent research in the selection of life insurance salesmen. *J. appl. Psychol.,* 1941, **25,** 11–17.

Kurtz, A. K. A research test of the Rorschach test. *Personnel Psychol.*, 1948, **1**, 41–51.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. *Psychol. Monogr.*, 1947, **61**, No. 285.

Maller, J. B. Personality tests. In J. McV. Hunt (Ed.), *Personality and the behavior disorders.* New York: Ronald, 1944.

Mandell, M. Testing for administrative and supervisory positions. *Educ. psychol. Measmt.*, 1945, **5**, 217–228.

Marks, E. S. Selective sampling in psychological research. *Psychol. Bull.*, 1947, **44**, 267–275.

McGehee, W. The prediction of differential achievement in a technological college. *J. appl. Psychol.*, 1943, **27**, 88–92.

McNemar, Q. Opinion—attitude methodology. *Psychol. Bull.*, 1946, **43**, 289–374.

McPherson, M. W. A method of objectively measuring shop performance. *J. appl. Psychol.*, 1945, **29**, 22–26.

Meehl, P. E. The dynamics of "structured" personality tests. *J. clin. Psychol.*, 1945, **1**, 296–303.

Meehl, P. E., and S. R. Hathaway. The K factor as a suppressor variable in the Minnesota multiphasic personality inventory. *J. appl. Psychol.*, 1946, **30**, 525–564.

Melton, A. W. *Apparatus tests.* Army Air Forces Aviation Psychology Research Report 4. Washington, D. C.: U. S. Government Printing Office, 1947.

Merrill, W. M. Sampling theory in item analysis. *Psychometrika*, 1937, **2**, 215–224.

Mosier, C. I. A note on item analysis and the criterion of internal consistency. *Psychometrika*, 1936, **1**, 275–282.

Mosier, C. I. On the reliability of a weighted composite. *Psychometrika*, 1943, **8**, 161–168.

Mosier, C. I. Rating of training and experience in public personnel selection. *Educ. psychol. Measmt.*, 1946, **6**, 313–329.

Mosier, C. I. A critical examination of the concepts of face validity. *Educ. psychol. Measmt.*, 1947, **7**, 191–205.

Mosier, C. I., M. C. Myers, and H. G. Price. Suggestions for the construction of multiple choice items. *Educ. psychol. Measmt.*, 1945, **5**, 261–271.

Munroe, R. L. Prediction of the adjustment and academic performance of college students by a modification of the Rorschach method. *Appl. psychol. Monogr.*, 1945, No. **7**, 1–104.

Munroe, R. L. Academic success and personal adjustment in college. In *Exploring individual differences.* Washington, D. C.: American Council on Education, Series 1, 1948, **12**, No. 32, 30–42.

Odell, C. W. The scoring of continuity or rearrangement tests. *J. educ. Psychol.*, 1944, **35**, 352–356.

Otis, J. L. The criterion. In H. W. Stead, C. L. Shartle, et al., *Occupational counseling techniques.* New York: American, 1940.

Owens, W. A. An empirical study of the relationship between item validity and internal consistency. *Educ. psychol. Measmt.*, 1947, **7**, 281–288.

Patterson, C. H. On the problem of the criterion in prediction studies. *J. consult. Psychol.*, 1946, **10**, 277–280.

Peters, C. C., and W. R. Van Voorhis. *Statistical procedures and their mathematical bases.* New York: McGraw-Hill, 1940.

Pockrass, J. H. Common fallacies in employee ratings. *Person. J.*, 1940, **18**, 262–267.

Ramsperger, A. G. *Philosophies of science.* New York: Crofts, 1942.

Rao, C. R. A statistical criterion to determine the group to which an individual belongs. *Nature, Lond.*, 1947, **160**, 835–836.

Reed, R. R. An empirical study in the reduction of the number of variables used in prediction, supplementary study C. In P. Horst, *The prediction of personal adjustment.* New York: Social Science Research Council Bulletin 48, 1941.

Richardson, M. W. The relation of difficulty to the differential validity of a test. *Psychometrika*, 1936*a*, **1**, 33–49.

Richardson, M. W. Notes on the rationale of item analysis. *Psychometrika*, 1936*b*, **1**, 69–75.

Richardson, M. W. Combination of measures, supplementary study D. In P. Horst, *Prediction of personal adjustment.* New York: Social Science Research Council Bulletin 48, 1941*a*.

Richardson, M. W. The logic of age scales. *Educ. psychol. Measmt.*, 1941*b*, **1**, 25–34.

Richardson, M. W. The interpretation of a test validity coefficient in terms of increased efficiency of a selected group of personnel. *Psychometrika*, 1944, **9**, 245–248.

Richardson, M. W. Selection of Army officers. In *New methods in applied psychology.* College Park: University of Maryland, 1947. Pp. 79–85.

Richardson, M. W., and G. F. Kuder. Making a rating scale that measures. *Person. J.*, 1933, **12**, 36–40.

Ruch F. L. The comparative efficiency of the multiple-cutting-score method and the Wherry-Doolittle method in selecting winch operators. OSRD, 1945 (PB 15820).

Ruch, F. L. A comparative study of the predictive efficiency of batteries of tests selected by the Wherry-Doolittle and a multiple-cutting-score method. (Abstract.) *Amer. Psychologist*, 1948, **3**, 291.

Rulon, P. J. A simplified procedure for determining the reliability of a test by split halves. *Harv. educ. Rev.*, 1939, **9**, 99–103.

Rulon, P. J. Validity of educational tests. In *National project in educational measurement.* American Council on Education, Series 1, 1947, **11**, No. 28, 13–20.

Rundquist, E. A. Development of an interview for selection purposes. In *New methods in ap-*

plied psychology. College Park: University of Maryland, 1947. Pp. 85–95.

Sarbin, T. R. The logic of prediction in psychology. *Psychol. Rev.,* 1944, **51,** 210–228.

Sarbin, T. R., and E. S. Bordin. New criteria for old. *Educ. psychol. Measmt.,* 1941, **1,** 173–186.

Segel, D. *Differential diagnosis of ability in school children.* Baltimore: Warwick and York, 1934.

Sells, S. B., and R. M. W. Travers. Observational methods of research. *Rev. educ. Res.,* 1945, **15,** 394–407.

Shartle C. L. *Occupational information.* New York: Prentice-Hall, 1946.

Shartle, C. L. Developments in occupational classification. *J. consult. Psychol.,* 1946, **10,** 81–92.

Shartle, C. L., B. J. Dvorak, C. A. Heinz, et al. Ten years of occupational research, 1934–1944. *Occupations,* 1944, **7,** 387–446.

Shipley, W. C., et al. The personal inventory — its derivation and validation. *J. clin. Psychol.,* 1946, **4,** 318–322.

Sisson, E. D. Forced Choice — the new Army rating. *Person. Psychol.,* 1949, **1,** 365–379.

Sletto, R. Construction of personality scales by the criterion of internal consistency. Hanover, N. H., and Minneapolis: Sociological Press, 1937.

Spence, K. The postulates and methods of behaviorism. *Psychol. Rev.,* 1948, **55,** 67–78.

Staff, Division of Occupational Analysis, War Manpower Commission. Factor analysis of occupational aptitude tests. *Educ. psychol. Measmt.,* 1945, **5,** 147–155.

Staff, Personnel Research Section, Personnel Research and Procedure Branch, The Adjutant General's Office. The forced choice technique and rating scales. (Abstract.) *Amer. Psychologist,* 1946, **1,** 267.

Stalnaker, J. M. Personnel placement in the armed forces. *J. appl. Psychol.,* 1945, **29,** 338–345.

Stalnaker, J. M., and M. W. Richardson. A note concerning the combination of test scores. *J. gen. Psychol.,* 1933, **8,** 460–463.

Stead, W. H., C. L. Shartle, et al. *Occupational counseling techniques.* New York: American, 1940.

Stevens, S. S. On the theory of scales of measurement. *Science,* 1946, **103,** 677–680.

Stewart, N. Relationship between military occupational speciality and army general classification test score. *Educ. psychol. Measmt.,* 1947, **7,** 677–693.

Strong, E. K. *Vocational interests of men and women.* Stanford: Stanford University Press, 1943.

Stuit, D. B. (Ed.). *Personnel research and test development in the Bureau of Naval Personnel.* Princeton: Princeton University Press, 1947*a*.

Stuit, D. B. The effect of the nature of the criterion upon the validity of aptitude tests. *Educ. psychol. Measmt.,* 1947*b*, **7,** 671–676.

Stuit, D. B., and J. T. Wilson. The effect of an increasingly well defined criterion on the prediction of success at Naval Training School (Tactical Radar). *J. appl. Psychol.,* 1946, **30,** 614–623.

Super, D. E. The validity of standard and custom-built personality inventories in a pilot selection program. *Educ. psychol. Measmt.,* 1947, **7,** 735–744.

Symonds, P. M. *Diagnosing personality and conduct.* New York: Century, 1931.

Taylor, H. C., and J. T. Russell. The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *J. appl. Psychol.,* 1939, **13,** 565–578.

Thomson, G. H. Weighting for battery reliability and prediction. *Brit. J. Psychol.,* 1940, **30,** 357–366.

Thorndike, E. L. The future of measurement of abilities. *Brit. J. educ. Psychol.,* 1948, **18,** 21–25.

Thorndike, R. L. *Research problems and techniques.* Army Air Forces Aviation Psychology Program Research Report 3. Washington, D. C.: U. S. Government Printing Office, 1947*a*.

Thorndike, R. L. Logical dilemmas in the estimation of reliability. In *National Project in Educational Measurement,* American Council on Education, Series 28, 1947*b*, **11,** 21–30.

Thorndike, R. L. *Personnel selection.* New York: Wiley, 1949.

Thurstone, L. L. *Multiple-factor analysis.* Chicago: University of Chicago Press, 1947.

Thurstone, L. L. Psychophysical methods. In T. G. Andrews (Ed.), *Methods of psychology.* New York: Wiley, 1948*a*.

Thurstone, L. L. Psychological implications of factor analysis. *Amer. Psychologist,* 1948*b*, **3,** 402–408.

Toops, H. A. The selection of graduate students. *Person. J.,* 1928, **6,** 470–471.

Toops, H. A. The successive hurdles method. *Person. J.,* 1932, **11,** 216–218.

Toops, H. A. The L-Method. *Psychometrika,* 1941, **6,** 249–266.

Toops, H. A. The criterion. *Educ. psychol. Measmt.,* 1944, **4,** 271–297.

Toops, H. A. Some concepts of job families and their importance in placement. *Educ. psychol. Measmt.,* 1945*a*, **5,** 195–216.

Toops, H. A. Philosophy and practice of personnel selection. *Educ. psychol. Measmt.,* 1945*b*, **5,** 95–124.

Travers, R. M. W. The use of the discriminant function in the treatment of psychological group differences. *Psychometrika,* 1939, **4,** 25–32.

Travers, R. M. W. A note on the value of customary measures of item validity. *J. appl. Psychol.,* 1942, **26,** 625–632.

Tucker, L. R. Maximum validity of a test with equivalent items. *Psychometrika,* 1946, **9,** 1–13.

Tucker, L. R. The problem of differential criteria. In *Exploring individual differences.* American Council on Education, Series 32, 1948, **12**, 30–42.

Turnbull, W. W. A normalized graphic method of item analysis. *J. educ. Psychol.,* 1946, **37**, 129–141.

VanDusen, A. C. Importance of criteria in selection and training. *Educ. psychol. Measmt.,* 1947, **7**, 498–504.

Vaughn, C. L. The nominating technique. In *New methods in applied psychology.* College Park: University of Maryland, 1947. Pp. 22–26.

Vernon, P. E. *The assessment of psychological qualities by verbal methods.* London: Medical Research Council Industrial Health Research Board Report 83, 1938.

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

Wallin, P. The prediction of individual behavior from case studies, supplementary study A. In P. Horst, *The prediction of personnel adjustment.* New York: Social Science Research Council Bulletin 48, 1941.

Warren, H. C. *Dictionary of psychology.* Boston: Houghton Mifflin, 1934.

Weider, A., K. Brodman, B. Mittelmann, D. Wechsler, and H. G. Wolff. Cornell service index: A method for quickly assaying personality and psychosomatic disturbances in men in the armed forces. *War med.,* 1945, **7**, 209–213.

Wexler, M. Measures of personal adjustment. In D. B. Stuit (Ed.), *Personnel research and test development in the Bureau of Naval Personnel.* Princeton: Princeton University Press, 1947.

Wherry, R. J. An approximation method for obtaining a maximized multiple criterion. *Psychometrika,* 1940, **5**, 109–115.

Wherry, R. J. Maximal weighting of qualitative data. *Psychometrika,* 1944, **9**, 263–266.

Wherry, R. J. Test selection and suppressor variables. *Psychometrika,* 1946, **11**, 239–247.

Wherry, R. J., and R. H. Gaylord. The concept of test and item reliability in relation to factor pattern. *Psychometrika,* 1943, **8**, 247–264.

Wherry, R. J., and R. H. Gaylord. Factor pattern of test items and tests as a function of the correlation coefficient: Content, difficulty and constant error factors. *Psychometrika,* 1944, **9**, 237–244.

Wherry, R. J., and R. H. Gaylord. Test selection with integral gross score weights. *Psychometrika,* 1946, **11**, 173–183.

Wherry, R. J., and E. K. Taylor. The relation of multiserial eta to other methods of correlation. *Psychometrika,* 1946, **11**, 155–161.

Wilks, S. S. Weighting systems for linear functions. *Psychometrika,* 1938, **3**, 23–40.

Williams, S. B., and H. J. Leavitt. Methods of selecting Marine Corps officer candidates. In *New methods in applied psychology.* College Park: University of Maryland, 1947. Pp. 96–99.

Wolf, R. R. Differential forecasts of achievement and their use in educational counselling. *Psychol. Monogr.,* 1939, **51**, No. 1.

Wolfle, D. L. Factor analysis in the study of personality. *J. abnorm. soc. Psychol.,* 1942, **37**, 393–397.

Zerga, J. E. Job analysis, a résumé and bibliography. *J. appl. Psychol.,* 1943, **27**, 249–267.