A Sampling Model for Validity

Michael T. Kane
National League for Nursing

A multifacet sampling model, based on generalizability theory, is developed for the measurement of dispositional attributes. Dispositions are defined in terms of universes of observations, and the value of the disposition is given by the universe score, the mean over the universe defining the disposition. Observed scores provide estimates of universe scores, and errors of measurement are introduced in order to maintain consistency in these estimates. The sampling model provides a straightforward interpretation of validity in terms of the accuracy of estimates of the universe scores, and of reliability in terms of the consistency among these estimates. A third property of measurements, import, is defined in terms of all of the implications of a measurement. The model provides the basis for a detailed analysis of standardization and of the systematic errors that standardization creates; for example, the hypothesis that increases in reliability may cause decreases in validity is easily derived from the model. The model also suggests an explicit mechanism for relating the refinement of measurement procedures to the development of laws and theories.

I. Introduction

The technical quality of behavioral measurements is evaluated in terms of two properties—reliability and validity. Validity involves the interpretation of the observed score as representative of some external property, and reliability deals with the consistency among observed scores. In general terms, reliability is concerned with precision and validity is concerned with accuracy (Stallings & Gillmore, 1971). Since a very precise estimate of the wrong attribute is less useful than a relatively imprecise estimate of the intended attribute, validity is generally considered to be more important than reliability. However, the evidence for the validity of most behavioral measurements is less adequate than the evidence for their reliability. Ebel (1961) has aptly described this dilemma:

> Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few. Test validation, in fact, is widely regarded as the least satisfactory aspect of test development. (p. 640)

This situation has not improved markedly since 1961. The statement often found in introductory textbooks equating validity with the extent to which scores measure "what they are intended to measure"

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 2, Spring 1982, pp. 125–160
© Copyright 1982 Applied Psychological Measurement Inc.
0146-6216/82/020125-36$2.80

provides an extreme example of the conceptual problems that surround validity, since it suggests the existence of a "true" value for an attribute, without specifying what this "true" value represents.

Ebel (1961) has also pointed out that physics does not seem to encounter problems of validation. Indeed, in the classic analysis of physical measurement, Campbell (1957) did not employ a separate concept of validity defined in terms of the relationship between observed values and "true" values. In developing and evaluating measurement procedures for physical properties, such as length and mass, the interpretation of the properties is closely tied to the observations that are used in their measurement, and therefore validity is built into the measurements. Because the connection between the interpretation of physical attributes and their measurement is often particularly straightforward, several of the examples used in this paper will involve physical measurement.

The multifacet sampling model developed in this paper is based on generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Although most of the results are stated in terms of variance components, the development of the model emphasizes conceptual rather than technical issues. The most useful results of the sampling model deal with the questions that should be asked and the general form of the answers that should be sought in analyzing measurements.

Measurements are analyzed by examining how observations are related to their intended interpretation. The model provides an analysis of reliability, validity, errors of measurement, the distinction between random errors and systematic errors, standardization, and the role of theory in interpreting measurement. Within the sampling model the issues of reliability, validity, and errors of measurement arise naturally as requirements for the intended interpretations to be meaningful. Since the development of the model's implications is quite long, and, in some ways relatively convoluted, an overview of the main points in the development may provide a useful road map.

Section II examines the interpretation of attributes as dispositions defined in terms of universes of possible observations. The value assigned to an attribute is defined as the expected value over this universe, and measurements are interpreted as estimates of this expected value. Because the estimates are based on samples from the universe defining the attribute, the model is a sampling model. Estimates of the attribute's value based on different samples will not generally be equal, and in order to maintain consistency, an explicit theory of errors is introduced.

Section III provides a brief outline of generalizability theory and introduces a sampling model for validity. The validity of measurements of a dispositional attribute is defined in terms of the accuracy with which the observed scores estimate the expected value for the appropriate universe.

Section IV examines the effects of standardization on measurement procedures. Standardized measurements involve two kinds of errors: random errors, which vary from one observation to another, and systematic errors, which are constant for a series of observations. Reliability is associated with random errors, and validity is associated with systematic errors.

Section V explores the relationship between theory and measurement. A third property of measurements is introduced by defining the concept of import in terms of all of the inferences that can be drawn from an observed score. Import is associated with the connotation of attribute labels, while validity is associated with denotation of attribute labels. This section reviews some powerful techniques for controlling errors of measurement.

Finally, Section VI examines the assumptions underlying the sampling model and presents some concluding comments. In particular, the problems associated with the sampling assumptions are discussed within the broader context of the problem of inductive inference.

II. The Interpretation of Measurable Attributes

Lord and Novick (1968, p. 17) define measurement as "a procedure for the assignment of numbers . . . to specified properties of experimental units in such a way as to characterize and preserve specified relationships in the behavioral domain." According to Nunnally (1967, p. 2), "Measurement consists of rules for assigning numbers to objects to represent quantities of attributes." Campbell (1957, p. 267) defines physical measurement as "the process of assigning numbers to represent qualities."

According to each of these definitions, measurement involves a functional relationship between real numbers and the members of some class of objects. Depending on the attribute being considered, the objects may take a variety of forms, including physical objects, persons, and various complex systems. The rules used to assign the numbers may also vary considerably. The process of measurement, however, always involves a mapping of object o into a real number, μ_o, representing the value of the attribute for o. The fact that the number is assigned to the object of measurement involves a fundamental theoretical commitment, in that it implies that the attribute depends only on the object of measurement and does not depend on any other conditions that may prevail when the observations are made. For example, the statement that the length of a metal bar is 1.5 meters treats length as a property of the bar and implies that length does not depend, for example, on the location, orientation, or temperature of the bar, or on the identity of the observer.

Attributes

There are at least three kinds of attributes in science (see Ellis, 1968): basic attributes, derived attributes, and theoretical attributes. This paper is concerned mainly with the measurement of basic and derived attributes, but theoretical attributes will also be discussed briefly under the heading of construct validity.

A *basic attribute* represents an observed ordering on some property. It is noticed, for example, that some objects are easier to move than others and that this ordering of the objects remains the same regardless of their location, who attempts to move them, or when they are moved. It is convenient, therefore, to think of "resistance to movement" as a property, or attribute, of the objects; and a large class of solid objects can be characterized by this property. Where such an ordinal property exists for a class of objects, a basic attribute can be developed by assigning numbers to the objects corresponding to their ordering. Basic attributes are generally the first kind of attribute developed in a science.

A basic attribute can always be viewed as a disposition, or a tendency, to produce a certain reaction to some test conditions (Carnap, 1953, 1966). Dispositions may be qualitative or quantitative. For a *qualitative disposition*, the object is said to have the attribute if and only if a specific reaction occurs in the presence of appropriate test conditions. For example, an object is said to be a magnet if, when placed near a small piece of iron that is free to move (the test condition), it causes the iron to move (the reaction). For a *quantitative disposition*, a number is assigned to the object on the basis of the strength of the reaction to the test conditions. The magnitude of the attribute of being magnetic, or the strength of a magnet, can be defined in terms of the force it exerts on a piece of iron.

The basic attribute, mass, is derived from the qualitative ordering of objects in terms of their resistance to movement; and the measurement of mass, using balances and springs, reflects this origin. In many cases, however, the procedures used to measure a basic attribute do not reflect the original qualitative ordering so closely. The attribute temperature is based on the ordering of objects in terms of perceived warmth or coolness. The ordinary operations for measuring temperature, however, involve the expansion of liquids. Since it is known empirically that the volume of a liquid is closely related to perceptions of warmth, and since measurements based on liquid thermometers have much higher interobserver agreement than perceptions of warmth, thermometers have been substituted for perceived warmth in measuring temperature.

Derived attributes are constants in empirical laws (see Meehl, 1950). After measurements of some basic attributes are available, empirical laws stating relationships among the basic attributes may be developed, and these laws often involve constants that can also be treated as measurable attributes. For example, the length of metal bars is found to depend on their temperature, and the empirical law stating this relationship contains a constant, *k*, called the coefficient of thermal expansion of the bar. The value of *k* varies from one bar to another but remains relatively constant from one observation to another on a given bar. It is convenient, therefore, to interpret *k* as a property of the bar by assuming that it depends on the bar but not on the conditions prevailing when the bar is observed. The constant *k* is called a derived attribute because its interpretation depends on, or is derived from, the law relating the basic attributes length and temperature.

Theoretical attributes are not directly observable but can be assigned a number indirectly because of their connection, through a theory, to one or more observable attributes. It is sometimes possible to explain a number of laws in terms of a few postulates defining a theory, and these postulates generally involve theoretical attributes. Through the network of relationships constituting the theory, the theoretical attributes are connected to the basic and derived attributes that appear in the laws that the theory is designed to explain. The existence of the theory makes it possible to interpret observed scores on two levels. First, they may be interpreted as more-or-less direct estimates of a basic or derived attribute. Second, the same observations may be interpreted as indirect indicators of an unobservable theoretical construct.

Operational Definitions

The rules that are used to assign a value to basic and derived attributes are usually called *operational definitions* (Bridgman, 1927). The rules are operational in the sense that they are stated in terms of the operations performed in measuring the attribute. The rules are said to be definitions because they provide an interpretation for the numbers assigned as values of the attribute (see Carnap, 1953; Ennis, 1973; Hempel, 1960).

Operational definitions generally include two kinds of rules—structural rules and selection rules. The *structural rules* specify the kind of observations that are to be used and the way in which numbers are derived from these observations. The structural rules may be more or less elaborate, but they always leave some issues open; for example, a particular observer would not be named. The *selection rules* specify the range of conditions that may be tolerated for the various characteristics of the observations. Some characteristics may be fixed, and some may be defined in terms of classes of conditions. It is assumed that the characteristics not mentioned in the structural rules need not be controlled at all.

Operational definitions do not specify particular observations; they specify classes, or universes, of observations. The definition of an attribute can always be made more precise and more complete by specifying particular conditions of observation, but it would be impossible to specify all of the characteristics that might influence an observation. Operational definitions are designed to achieve some generality of application while providing a clear indication of the kind of observations allowed. The universe of observations specified by the definition is generally somewhat fuzzy, in the sense that there are marginal cases in which it is not clear whether the observation should be included in the universe. This ambiguity is tolerated because it facilitates the development of scientific laws (Toulmin,

1953) and does not represent a serious problem in practice. It does, however, complicate the interpretation of sampling models, and these complications will be examined in the last section of this paper.

It is sometimes claimed that measurements of operationally defined attributes are valid by definition. It is maintained that the operations used to measure the attribute define the attribute, and the results of these operations are, by definition, the values of the attribute. According to this view, no interpretation is to be given to measurements beyond the fact that they result from particular operations. In practice, however, the operational definitions of even the most narrowly defined attributes involve classes of observations rather than particular observations. No operational definition in science specifies a particular observer (John Jones), particular equipment (voltmeter No. 6), or a particular time and place. If the results of a particular observation could not be interpreted in terms of a universe of similar observations, these results would be of little interest. To assign a value to an attribute is to make a claim about a universe of observations.

Although restrictions may be placed on the qualifications of observers and on the type of equipment used, these restrictions define classes of observations rather than particular observations. A complete description of a single observation would require exhaustive specification of all of the conditions under which the observation is made. The definition of a dispositional attribute specifies only some of the conditions of observation, thereby allowing the other conditions to vary, and describes a class of observations rather than a single observation. Since few of these observations will actually be made for any object of measurement, the value of the attribute is inferred rather than observed. Therefore, each observation provides information about a universe of observations that could have been made. This generalization from particular observations to the universe defining an attribute is a cardinal feature of measurement.

The Object of Measurement

The object, or unit, to which a number is assigned by measurement is the *object of measurement*. The operational definition of an attribute specifies a class of observations for each object of measurement, and any of these observations could be used to estimate the value of the attribute for the object of measurement. Also, each observation can provide information about different objects of measurement; and if the measurement is to be interpreted unambiguously, the object of measurement must be clearly identified.

For example, in a study of anxiety, an observation might consist of the response of a person to some stimulus in a particular context. For such observations, the person is usually considered to be the object of measurement, and the level of anxiety is attributed to the person. In examining the degree to which various stimuli or contexts provoke anxiety, however, the objects of measurement would be stimuli or contexts, respectively. More complicated objects of measurement can also be considered. For example, the differential impact of stimuli on different persons could be described by taking person-stimulus pairs as the objects of measurement; in this case, the attribute would indicate how much anxiety the stimulus causes in the person. Cardinet, Tourneur, and Allal (1976) have discussed how the interpretation of an observation depends on the definition of the object of measurement.

The specification of the object of measurement is a conceptual issue and is not uniquely determined by the nature of the observations that are made. As the above examples illustrate, a single observation can provide information about a variety of objects of measurement. Similarly, many different observations may be used to measure a particular attribute for an object of measurement. The set of all possible objects of measurement for an attribute will be referred to as the *population* for the attribute.

The distinction that is often drawn in psychology between a state and a trait depends on a distinction between different kinds of objects of measurement. If the object of measurement is considered to be a person in a particular context, then the attribute being measured is a state variable, which is assumed to be a function of both the person and the time. It is expected that the value associated with a state variable will change as the context changes over time. For a trait, the object of measurement is the person, and the value of the trait variable is assumed to be independent of time. It is recognized, of course, that the behaviors associated with the trait may be exhibited to different degrees in different contexts, but the value assigned to the trait is assumed not to change. For a trait variable, changes in the observed variable over time are treated as errors of measurement; for a state variable such differences are accounted for by differences in the value of the state variable.

In the physical sciences, distinctions among different kinds of objects of measurement are often drawn very carefully. In their introductory treatment of mechanics, Corben and Stehl (1960) state the following assumptions:

> A particle is described when its position in space is given and when the values of certain parameters such as mass, electric charge, and magnetic moment are given. By our definition of a particle, these parameters must have constant values because they describe the internal constitution of the particle. If these parameters do vary with time, we are not dealing with a simple particle. The position of a particle may, of course, vary with time. (p. 6)

Therefore, the mass, charge, and magnetic moment are to be treated as trait variables, with particles as their objects of measurement; position is to be treated as a state variable which varies over time and therefore has particle-time combinations as its objects of measurement.

The Use of Invariance Properties as Inference Tickets

Attributes are "constructed" by specifying universes of observations. Measurements of attributes are based on samples from these universes. In order to interpret a measurement as the value of an attribute for the object of measurement, there must be generalization from a sample of observations to a universe of observations. A central concern of a theory of measurement is therefore the justification of such inferences.

The evidence for a scientific inference is generally provided by appeal to laws (Hempel, 1965); and because the justification of inferences is their major function, Toulmin (1953) refers to scientific laws as "inference tickets." The type of law that is needed to justify the interpretation of observations as measurements is an invariance property. An *invariance property*, or invariance law, states that the results of a certain kind of observation do not depend on some of the conditions of observation. Invariance laws are needed for the interpretation of measurement because measurement assigns a value to an object of measurement and not to an observation.

The attribute is identified with a universe of observations and not with a particular observation. Any observation from this universe could be used to assign a value for the attribute to the object of measurement. The *observed score*, X_{oi}, for an observation is the real number assigned to the observation by the structural rules for the attribute. A different but equally legitimate observed score, $X_{oi'}$, could be obtained by changing the conditions of observation in accordance with the selection rules.

When an observed score is interpreted as a measurement, it is assumed that this observed score represents the value of the attribute for the object of measurement. Since different observations will yield different observed scores for an object of measurement, the following relationship must hold, at least approximately, in order to avoid inconsistency:

$$X_{oi} = X_{oi'}, \qquad [1]$$

where i and i' represent any two observations that satisfy the definition of the attribute. That is, the observed scores must be approximately invariant over the universe of observations defining the attribute. Since the two quantities in Equation 1 are observable, this assertion is testable for any pair of observations, and Equation 1 is an empirical hypothesis.

Equation 1 is an invariance law stating that the observed scores are invariant over the universe defining the attribute. For a given object of measurement, o, all observations included in the definition of the attribute should assign the same value to the object of measurement; and if they do, this common value may be taken to be the value of the attribute for the object of measurement, o. Note that the invariance properties required for the measurement depend on the definition of the attribute being measured. If the two quantities in Equation 1 were not taken as measurements of the same attribute for the same object of measurement, there would be no reason to require that they should have the same value.

Considering again the example discussed earlier, if anxiety is interpreted as a trait, the situations in which observations are made are conditions of observation, and invariance over situations is assumed. If anxiety is interpreted as a state, the objects of measurement are persons in situations, and changes in the attribute value as a function of the situation are consistent with this interpretation.

Invariance laws are involved in measurement because they justify inferences from samples of observations to a universe of observations. If all of the observations in the universe give the same result for any object of measurement, then any one of these observations would provide complete information about the universe. If Equation 1 holds for all pairs of observations defining an attribute, it provides the necessary justification for inferences from observed scores to the attribute value. To the extent that observations fail to satisfy Equation 1, such inferences will be inconsistent (Suppes, 1974). Therefore, the invariance law in Equation 1 is necessary for the interpretation of observations as measurements of dispositional attributes.

Errors of Measurement

The inconsistency arising from violations of Equation 1 can be eliminated by introducing the concept of an error of measurement. The result of any observation on an object, o, is taken to be the sum of the "true" value of the attribute, t_o, plus an error of measurement, e_{oi},

$$X_{oi} = t_o + e_{oi}. \qquad [2]$$

Since neither the "true" score nor the error is directly observable, Equation 2 is not a testable hypothesis; rather, it is a definition of the error, e_{oi}.

However, the values assigned to the errors are not arbitrary. Given the universe defining an attribute and a value for the true score, the errors are determined empirically. If the observed scores have approximately the same value, the values assigned to the errors can be small. If the observed scores on a given object of measurement vary widely, the values assigned to the errors must be large. Small errors of measurement are, of course, generally perferred over large errors of measurement, and therefore Equation 2 provides a relative criterion for evaluating measurement procedures.

Classical test theory defines the value of an attribute for an object of measurement as the expected value over all observations included in the definition of the attribute. This choice is convenient because it minimizes the mean-square error. With this definition of true score, it is easy to show that the expected value of the errors is zero for each object of measurement and, therefore, that the error variance for each object of measurement is equal to its observed score variance. Such object-specific error variances are very useful because they indicate the accuracy of estimates of the true score for each object of measurement. However, the direct estimation of the object-specific error variance requires repeated observations on each object of measurement, and this is often not practical.

A more easily estimated parameter is the average error variance, $\sigma^2(e)$, over all objects of measurement. The average error variance is more widely used than the object-specific error variance because it can be estimated with pairs of observations on each object of measurement. The covariance between true scores and errors of measurement can be shown to be zero, and therefore the observed score variance can be partitioned as

$$\sigma^2(X) = \sigma^2(t) + \sigma^2(e), \qquad [3]$$

where $\sigma^2(t)$ is the variance in the true scores over the population and $\sigma^2(e)$ is the average error variance.

Errors of Measurement as Constructs

In the absence of assumptions about attributes and objects, the concept of an error of measurement is unnecessary. If attention is restricted to observations, there is no reason to reject the hypothesis that every observation is perfectly accurate. Suppose, for example, that two observers put thermometers into the same glass of water at the same time. Suppose further that one of the observers records the temperature as 20° C and the other observer records the temperature as 22° C. These two observations differ in several ways, for example, in terms of the observer, the thermometer, and the position in the water. If the two numbers, 20 and 22, are assigned to the observations, there is no reason to assume that either observation should be said to contain any error. The two observations occurred as they occurred. The need for a concept of error arises only when attention is shifted from observation to measurement and assumptions about invariance are introduced.

The usual analysis of the example given above considers temperature to be the attribute and the glass of water at a particular time to be the object of measurement. The temperature is assumed to be a function of the water and the time and to be invariant over thermometers, locations in the glass, and observers. This implies that the two observations described above should agree with each other, and any discrepancy between them is explained by an error of measurement.

In general, any two observations on an object of measurement will produce different numerical results. Since measurement is intended to map each object into one real number, theory must be adjusted in one of two ways. One approach is to redefine the objects of measurement so that the different observations apply to different objects. In the example above, the objects of measurement could be redefined to be the small volumes of water surrounding each thermometer, thus explaining the differences between the two measurements by the fact that they apply to different objects. An alternative approach leaves the definition of the object of measurement unchanged but introduces an explicit theory of errors. It is thereby recognized that the observations used in measurement depend on the conditions of observation and not just on the object of measurement.

Relative Error

For many applications, the error variance is not a very good index for the accuracy of measurement. The magnitude of the error variance could be changed simply by changing the scale (e.g., inches to feet), and the evaluation of a measurement procedure should not depend on such an arbitrary choice. It is not the absolute magnitude of the error variance that is significant but the magnitude relative to the degree of precision needed for some purpose.

The precision required of measurement procedures varies widely. An astronomer can often tolerate crrors of thousands of kilometers, while a crystallographer might consider an error of a thousandth of a centimeter to be unacceptable. The magnitude of the errors that can be tolerated depends on the magnitude of the quantities being measured, and therefore the degree of precision is often reported relative to the magnitude of the quantities being measured. The practice of reporting the magnitude of measurement errors in relative terms is general enough in the physical sciences to be included in an introductory textbook (Physical Science Study Committee, 1968):

> If a surveyor measures a distance with great care he might get 100.132 meters ±0.3 cm. His work is a great deal more accurate than that done when the width of a book page is measured to the nearest millimeter with a ruler, even though his error is something like three times as big as what anyone would perhaps make on the page in ten seconds' work. This sometimes finds expression in another way when the estimated spread of measurements, the *tolerance*, is stated, using decimal fractions, or percentage. Thus the surveyor would say his length was 100.132 meters ±0.003%, while the page is just 20.1 cm. ±0.5%. (p. 14)

The emphasis on stating the magnitude of the errors in relative terms has been even more pronounced in the social sciences (Lord & Novick, 1968):

> . . . the effectiveness of a test as a measuring instrument usually does not depend merely on the standard error of measurement, but rather on the ratio of the standard error of measurement to the standard deviation of observed scores in the group. (p. 252)

A suitable index for the relative magnitude of errors of measurement is suggested by the fact that measurements are based on qualitative orderings of some kind, and the numbers assigned by the measurement procedure should reflect this qualitative ordering. As a minimal requirement, the errors should not be so large as to cause significant fluctuations in the ranks assigned to objects from one set of observations to another (Cronbach & Gleser, 1964). The consistency of the ranking of objects of measurement from one set of observations to another can be estimated by the correlation between the two sets of observed scores. Correlations indicate the degree of linear relationship between two variables; but in the absence of serious departures from linearity, they reflect the consistency of rankings from one variable to the other. Therefore, correlation coefficients and indices that are closely related to correlation coefficients (i.e., generalizability coefficients) have been widely used in evaluating the precision of measurements. In particular, correlation coefficients constitute the basic mathematical machinery in classical test theory.

The Role of Theory

The analysis of measurement errors depends on the assumption that attributes apply to specific kinds of objects of measurement and that certain invariance properties hold. The introduction of an explicit theory of errors represents a decision not to study some kinds of phenomena. In the example discussed earlier, the decision to interpret the difference between the two thermometer readings, 20 and 22, in terms of errors of measurement is essentially a decision not to investigate temperature variations within the liquid; this decision, which is not dictated by empirical findings, reflects a choice among several possible research strategies. By specifying the objects to be studied and the attributes to be assigned to these objects, the interpretation given to measurements shapes and is shaped by theory.

In order to make its task more manageable, every science tends to restrict what it treats explicitly. Errors of measurement provide a way of handling observed variations that are not to be given an explicit description or explanation. This makes it possible to minimize the number of objects of measurement that need to be considered and therefore to simplify both descriptions of phenomena and the theories designed to explain phenomena. As the science develops, it may be able to analyze phenomena that had earlier been relegated to error, thus decreasing the error; but there is always some variation which is intentionally left unexplained. Errors of measurement may be viewed as concessions to the brute fact that the world of observations is not as simple as might be desired.

The specification of the attributes and the objects of measurement to be studied determines how observations are described and organized, and this influences the kinds of questions addressed by the science, i.e., the paradigm for the science. A change in the definitions of attributes and objects of measurement (i.e., a change in the definition of error) represents a shift in the way that phenomena are perceived and described, and the resulting changes may be significant enough to be called a scientific revolution (Kuhn, 1970). For example, the changes introduced into physics by the special theory of relativity are basically changes in the invariance properties associated with length and time (Frank, 1953). In classical mechanics, length and time are assumed to be invariant with respect to the observer; but in the theory of relativity, length and time depend on the observer's frame of reference. The special theory of relativity had a revolutionary impact because it modified the objects of measurement, and therefore the assumed invariance laws, for the fundamental attributes of physics.

III. A Sampling Model for Validity

By definition, the value of an attribute for an object of measurement is the expected value over all observations in the universe defining the attribute for the object. If this universe score were available, it would be a perfectly accurate measure of the dispositional attribute. However, the universe score is generally not available, and samples of observations must be used to estimate it. This suggests the following definition of *validity* for attributes which are interpreted as dispositions:

> A measurement procedure is said to be valid for a dispositional attribute to the extent that it provides accurate estimates of the expected value over the universe of observations defining the attribute.

Validity reflects the accuracy of inferences from an observed score to the value of the attribute—the expected value over the universe—where accuracy is defined in terms of the expected squared error in estimation. Validity is a matter of degree, rather than all or none, and depends on the design of the measurement procedure and the interpretation of the attribute.

The sampling model based on this definition of validity is a multifacet model in the sense that the universe defining an attribute may involve observations that vary along a number of dimensions or facets. Previous discussions of sampling models for validity (e.g., see Kaiser & Michael, 1975; McDonald, 1978; Tryon, 1957) have generally been restricted to sampling from a single facet. By employing generalizability theory (Cronbach et al., 1972), the multifacet sampling model discussed in this paper provides a more comprehensive analysis of the sampling designs associated with measurement procedures than the unifacet sampling models can provide.

Of the many approaches to validity that have been suggested (Cronbach, 1971), construct validity is the most general and can be interpreted as including all of the others. It emphasizes the legitimacy with which various inferences can be drawn on the basis of observed scores and allows for a wide range of techniques, corresponding to the range of inferences to be drawn. In its emphasis on inferences from observed scores to the expected value over a universe of observations, the sampling model can be interpreted as a type of construct validity. The sampling model also raises questions usually included under content validity. The specification of the universe is, of course, a central concern for the sampling model, and this involves the kind of issue that is treated by content validity. Criterion validity is not included in this discussion of the sampling model because it does not come into play until after a valid criterion is available; the sampling model provides a mechanism for developing such criteria for dispositional attributes.

The remainder of this paper develops the implications of the sampling model. In this section, it is assumed that measurements consist of random samples from the universe defining the attribute being measured; this assumption is relatively unrealistic for many measurements, but the analysis of this simple case provides a convenient vehicle for elaborating the definition of validity and for developing some notation. This section also provides a brief introduction to generalizability theory.

Generalizability Theory

Generalizability theory (see Brennan, in press; Cronbach et al., 1972) allows for the existence of multiple sources of variation in measurements and uses ANOVA to estimate variance components for different effects. An observation on an object of measurement is assumed to be sampled from a universe of observations. The observations in the universe are described by the conditions under which they are made, and the set of all conditions of a particular type is called a *facet.* For example, in behavioral measurement, the universe often includes an item facet, an occasion facet, and a rater facet.

Cronbach et al. (1972, p. 20) have drawn a distinction between *G studies,* or generalizability studies, which estimate the variability associated with various facets, and *D studies,* or decision studies, which provide the data for substantive decisions. The purpose of the G study is to estimate components of variance, which may then be used to evaluate the dependability of measurement. In this paper, the term "measurement procedure" will often be used in place of the term "D study." A *measurement procedure* employs the same design over a number of separate studies. The term "D study" suggests that the sampling design for measurements of an attribute is likely to change from one study to another. Although the possibility of such changes in design is explicitly considered at several places in this paper, much of the discussion will focus on standardized procedures which may be used in several D studies.

Based on the distinction between G studies and D studies, Cronbach et al. (1972, p. 20) distinguish between two universes. In conducting a G study, certain facets are investigated and variance components for these facets are estimated. The facets investigated in the G study define a *universe of admissible observations.* In interpreting the observations in a D study as measurements, inferences are drawn to the universe of observations defining an attribute, and this universe is called the *universe of generalization.*

The *universe score* is the expected value of the observed score over the universe of generalization. Universe scores are not directly observable but can be estimated by the mean over a sample of observations; that is, for each object of measurement, the observed score is used as an estimate of the universe score. Therefore, generalizations from observed scores to universe scores are of central concern, and the dependability of such generalizations is described by a generalizability coefficient.

Cronbach et al. (1972, p. 97) define a *generalizability coefficient* as the ratio of the universe score variance to the expected observed score variance for the D study design. The universe score variance in a generalizability coefficient is analogous to the true score variance of classical test theory, and the expected observed score variance is analogous to the observed score variance of classical test theory. A generalizability coefficient can be interpreted in two ways (Kane & Brennan, 1977). First, it is approximately equal to the correlation between observed scores for two independent random samples of observations from the universe of generalization. Second, it is approximately equal to the expected value of the squared correlation between the observed score and the universe score.

A Linear Model

Generalizability theory allows for the use of a variety of linear models in designing and interpreting both G studies and D studies. The universe of generalization typically involves a number of facets; and, in principle, the model for observed scores could explicitly represent each of these facets. For the sake of simplicity, however, a one-facet model with replications will be used as a basis for discussion throughout this paper. In this simple model, only one facet is considered explicitly; all other facets are assumed to be sampled randomly and independently and are subsumed under the replication facet. The observed scores are represented by the linear model:

$$X_{oir} = \mu + \alpha_o + \alpha_i + \alpha_{oi} + \alpha_r, \qquad [4]$$

where

μ is the grand mean;
α_o is the main effect for the object of measurement, o;
α_i is the main effect for the i facet;
α_{oi} is the oi interaction; and
α_r is the replication effect.

The linear model in Equation 4 represents the observed scores in the universe of generalization; it is not intended to represent the sampling design for any particular G study or D study. Because the effects are defined in terms of differences among expected values of observed scores (e.g., α_o is defined as the expected observed score for the object, o, over i and r minus the grand mean, μ), each effect in the model is uncorrelated with every other effect, and the expected value of each effect over any of its subscripts is zero. Equation 4 is essentially a generalization of Equation 2. The main difference between the classical test theory model in Equation 2 and the linear model in Equation 4 is that the classical model assumes the existence of only two sources of variance in the observed score, while the model in Equation 4 explicitly considers four sources of variance and could easily be extended to include additional facets.

The model in Equation 4 includes two facets, labeled i and r; and for each of these facets, there is a universe of conditions from which the conditions in a particular study may be drawn. The universe for each facet may be either finite or infinite, but for the sake of simplicity, it is assumed in this paper that the universe of conditions for each facet is infinite. From a G study in which the i facet is crossed with objects of measurement, o, and replications are nested within oi combinations, four components of variance can be independently estimated. The variance components for the four random effects in Equation 4 are designated as $\sigma^2(o)$, $\sigma^2(i)$, $\sigma^2(oi)$, and $\sigma^2(r)$.

In a D study, the observed scores are usually based on the sum or average taken over a sample of observations, and capital letters are used to designate the average effect over a sample of observations. The variance components for the average values of the i main effect, the oi interaction, and replications, over a sample of n_i conditions of the i facet and n_r replications, are given by

$$\sigma^2(I) = \sigma^2(i)/n_i, \qquad [5a]$$

$$\sigma^2(oI) = \sigma^2(oi)/n_i, \qquad [5b]$$

$$\sigma^2(R) = \sigma^2(r)/n_i n_r. \qquad [5c]$$

The relationships listed in Equations 5a to 5c can be used to estimate variance components for D studies involving any values for n_i and n_r once the required random effects variance components are estimated in G studies. The estimation of variance components is discussed in detail by Cornfield and Tukey (1956), Cronbach et al. (1972), Lindquist (1953), Brennan (1977, in press), Smith (1978), and by textbooks on experimental design (e.g., Winer, 1971). Some of the virtues and limitations of generalizability theory are discussed by Rozeboom (1966), Ebel (1974), and Lumsden (1976).

Measurements Based on Random Sampling from the Universe of Generalization

Most measurement procedures do not involve independent random sampling from the universe of generalization, but it is convenient to start with this assumption. For a D study with *i* nested within *o* (a separate sample of conditions of the *i* facet is drawn for each observation on each object of measurement), the observed scores can be represented as

$$X_{oIR} = \mu + \alpha_o + \alpha_I + \alpha_{oI} + \alpha_R \qquad [6]$$

where *o* represents the object of measurement, *I* indicates a sample of n_i conditions of the *i* facet, and *R* indicates a sample of n_r replications for each condition of the *i* facet. Again, the replication index represents the effect of all facets other than the *i* facet. Since the effects in Equation 6 are uncorrelated, the expected observed score variance over the population and over the universe of generalization is

$$\sigma^2(X) = \sigma^2(o) + \sigma^2(I) + \sigma^2(oI) + \sigma^2(R). \qquad [7]$$

The universe score μ_o, for the object of measurement, *o*, is given by

$$\mu_o = \underset{IR}{\xi\xi}(X_{oIR}) = \mu + \alpha_o, \qquad [8]$$

and the universe score variance is given by

$$\underset{o}{\xi}(\mu_o - \mu)^2 = \sigma^2(o). \qquad [9]$$

Where observations are randomly sampled from the universe of generalization for each object of measurement, the expected value of the observed score over repeated applications of the measurement procedure is equal to the universe score, and the observed score is an unbiased estimate of the universe score.

In analyzing errors of measurement, Cronbach et al. (1972, p. 76) distinguish between the error Δ in point estimates of universe scores and the error δ in estimates of the universe score expressed as deviations from the grand mean, μ. The error of measurement for a point estimate of μ_o, based on X_{oIR}, is

$$\Delta_{oIR} = X_{oIR} - \mu_o = \alpha_I + \alpha_{oI} + \alpha_R. \qquad [10]$$

Since *I* and *R* are randomly sampled for each measurement, the expected value of Δ_{oIR} over repeated measurements is zero, and the observed score is an unbiased estimate of the universe score for *o*. The expected value of the squared error in point estimates, over *I* and *R*, is given by

$$\underset{IR}{\xi\xi}(\Delta^2_{oIR}) = \sigma^2(I) + \sigma^2(oI) + \sigma^2(R), \qquad [11]$$

which represents the error variance for point estimates of universe scores.

If conditions of the i facet are sampled independently for each observation, the expected value of X_{oIR} over the population is equal to the grand mean, μ, and the error in estimating universe deviation scores is

$$\delta_{oIR} = (X_{oIR} - \mu) - (\mu_o - \mu) = \alpha_I + \alpha_{oI} + \alpha_R. \qquad [12]$$

Since I and R are independently sampled for each observation of o, the expected value of δ_{oIR} over the universe is also zero, and the observed deviation score is an unbiased estimate of the universe deviation score. The variance in δ_{oIR} is equal to the variance in Δ_{oIR}, as given in Equation 11.

The covariance, taken over the population, of the errors Δ_{oIR} on two administrations of the measurement procedure is given by

$$\text{cov}(\Delta_{oIR}, \Delta_{oI'R'}) = \underset{o}{\xi}(\alpha_I + \alpha_{oI} + \alpha_R)(\alpha_{I'} + \alpha_{oI'} + \alpha_{R'}). \qquad [13]$$

Since the i and r facets are sampled independently for each observation, taking an expected value over o automatically involves taking expected values over I, I', R, and R'. Therefore, the expected value over each of the crossproducts in Equation 13 is zero, and the errors Δ_{oIR} are uncorrelated. Similarly, the covariance of the errors δ_{oIR} for two administrations of the measurement procedure is also equal to zero.

In classical test theory, errors of measurement have an expected value of zero and are uncorrelated across pairs of observations. Errors of measurement that satisfy these requirements will be called *random errors*. It is clear from the discussion just presented that for a measurement procedure based on independent random samples from the universe of generalization, all errors of measurement are random errors.

Cronbach et al. (1972) define a generalizability coefficient as the ratio of universe score variance to expected observed score variance. From Equations 7 and 9, the generalizability coefficient is given by

$$\xi\rho^2(X_{oIR}, \mu_o) = \frac{\sigma^2(o)}{\sigma^2(o) + \sigma^2(oI) + \sigma^2(I) + \sigma^2(R)}, \qquad [14]$$

where the notation "$\xi\varrho^2$" emphasizes the interpretation of the coefficient as an index of the squared correlation between observed scores and universe scores. The generalizability coefficient in Equation 14 incorporates tests of two separate invariance laws—one for the i facet and the other for the replication facet. Since the replication facet represents the effects of all but one of the facets in the universe of generalization, the second of the two invariance laws is very general. If the observed scores for each object of measurement are approximately invariant over the i facet, the variance components for the i effect and the oi interaction will be small. Similarly, if observed scores are invariant over all other facets in the universe of generalization, the replication variance component will be small. In general, each of the variance components in the error variance is associated with an invariance law.

The value of the generalizability coefficient depends on how thoroughly the measurement procedure samples the universe of generalization, and this is determined by the design of the procedure and by the definition of the attribute being measured. In particular, the more narrowly the universe of generalization is conceived, the more dependable the measurements will be.

By using Equation 5 with Equation 14, the generalizability coefficient can be estimated for any number of conditions of the i facet and any number of replications. Increasing the sample sizes for various facets provides a simple way of improving the dependability of measurements. However, there are practical limits on sample sizes; and later in this paper, more sophisticated approaches to the control of measurement errors will be discussed.

A Universe Sampling Model for Validity

Since the universe score for each object of measurement has been stipulated to be the value of the attribute for the object, a measurement procedure is valid to the extent that it estimates the universe scores accurately. For a measurement procedure consisting of random sampling from the universe of generalization, the observed score is an unbiased estimate of the universe score, and the random errors in Equation 10 are the only errors of measurement. Since the generalizability coefficient in Equation 14 indicates how accurately universe scores can be inferred from observed scores, it can be interpreted as a validity coefficient. Therefore, if a dispositional attribute were clearly specified in terms of a universe of generalization and if random samples could be drawn from this universe, validation would be relatively straightforward. Unfortunately, the universe of generalization is usually not so clearly defined, and this complicates the analysis of validity.

Although a generalizability coefficient can be an index of validity, most estimated generalizability coefficients are not validity coefficients. The interpretation of Equation 14 as a validity coefficient depends on the strong sampling assumption that the observed scores are based on random samples from the intended universe of generalization. For most observed scores, inferences are made to a universe of generalization that is much broader than the universe from which the observations are sampled. It is not unusual, for example, for inferences to be drawn about broadly defined universes of behaviors on the basis of responses to a particular type of written test item. In such cases, it is unreasonable to assume that the observations are a random sample from the universe of generalization for the attribute. Cronbach et al. (1972, p. 352) have pointed out that "investigators often choose procedures for evaluating the reliability that implicitly define a universe narrower than their substantive theory calls for. When they do so, they underestimate the 'error' of measurement, that is, the error of generalization."

For most attributes, standardization is used to control errors of measurement, which tend to be unacceptably large when observations are randomly sampled from the universe of generalization, and standardization involves an explicit decision not to use random sampling. A standardized measurement procedure samples observations from a subuniverse of the universe of generalization and therefore requires a more elaborate model for validity than that presented in this section.

Also, it is typically the case that there are unintended violations of the sampling assumptions in the G study. The effects of unintended departures from the random sampling assumption cannot be evaluated accurately, and therefore the interpretation of G-study results must always be somewhat tentative. The violation of sampling assumptions is, of course, a general problem in research, and the clouding of interpretations that results from such violations is not unique to the sampling model or to generalizability theory. However, sampling problems tend to be especially acute in G studies because the number of variance components to be estimated may be quite large. Establishing the validity of a measurement procedure requires the empirical testing of a number of invariance laws, and this task is not necessarily a simpler task than the testing of other empirical laws. The problem of induction that arises in verifying scientific laws and some of the solutions that have been proposed will be discussed more fully in the last section of this paper.

IV. Standardization and the Universe of Allowable Observations

As indicated earlier, the inclusion of an explicit theory of errors makes it possible for relatively simple theories to provide a consistent account of a wide range of observations. The inconsistency that would otherwise arise in the theories because of violations of invariance properties is accounted for by the errors of measurement, and since the magnitude of these errors can be estimated, their effects can be taken into account in interpreting the results of measurement. Although an explicit theory of errors is always useful, its advantages are most pronounced when the errors involved are small. It is desirable that the error variance be as small as possible because errors decrease the accuracy of the inferences that can be drawn from measurements.

There are three ways to decrease the error variance and therefore to increase the precision of measurement. The first way is to base each measurement on a larger sample of observations; this approach is widely used in both the physical and behavioral sciences and is discussed in detail by Cronbach et al. (1972). A second way to reduce errors is to change the definition of the attribute by restricting the universe of generalization; the more narrowly the universe of generalization is defined, the smaller the errors will be. The third method for controlling errors of measurement is to standardize some aspects of the measurement procedure. Standardization can be very effective in reducing errors of measurement, but it can also be misleading and therefore requires careful examination. The remainder of this section deals with standardization.

Standardization of Measurement Procedures

Since errors of measurement result from variations in the conditions of observation, the errors may be reduced by controlling the conditions of observation. If the observations on an object of measurement vary as some facet varies, these observations may be made more consistent by having all observations involve the same condition of the facet.

If all applications of a measurement procedure employ a particular condition, or set of conditions, of a facet, the procedure is said to be *standardized* on the facet. Standardization of the *i* facet changes the design of the measurement procedure so that all observations are associated with the same conditions, I^*, of the *i* facet, but it does not alter the definition of the attribute. Standardization is not intended to imply a change in the universe of generalization, and the universe score for object, o, is still μ_o, as given by Equation 8.

The observed score for a measurement procedure with the *i* facet standardized to I^* can be represented by

$$X_{oI^*R} = \mu + \alpha_o + \alpha_{I^*} + \alpha_{oI^*} + \alpha_R . \quad [15]$$

The expected value of the observed score over repeated application of the standardized measurement procedure is given by

$$\mu_o^* = \underset{R}{\xi}(X_{oI^*R}) = \mu + \alpha_o + \alpha_{I^*} + \alpha_{oI^*} . \quad [16]$$

For a standardized measurement procedure, therefore, the observed score is a biased estimate of the universe score unless the last two terms in Equation 16 happen to be zero. The error for point estimates of universe scores also reflects this bias:

$$\Delta_{oI^*R} = X_{oI^*R} - \mu_o = \alpha_{I^*} + \alpha_{oI^*} + \alpha_R . \quad [17]$$

Equation 17 differs from the corresponding expression for the error of an unstandardized measurement procedure, as given by Equation 10, in that its first term is a constant for all observations and its second term is a constant for all observations on a particular object of measurement. The expected value of the error Δ_{oI^*R} over repeated observations on the object, o, is given by

$$\underset{R}{\xi}(\Delta_{oI^*R}) = \alpha_{I^*} + \alpha_{oI^*} . \quad [18]$$

The two constants in Equation 18 represent bias in the standardized procedure's estimates of universe scores. The expected squared error over replications for object, *o*, is given by

$$\underset{R}{\xi} (\Delta^2_{oI*R}) = (\alpha_{I*} + \alpha_{oI*})^2 + \sigma^2(R). \qquad [19]$$

The expected value of Equation 19 over random samples, *I**, from the *i* facet is the same as the expected value of the corresponding squared error Δ^2_{oIR} for the unstandardized procedure, given by Equation 11. Therefore, standardization on randomly chosen conditions of a facet does not generally decrease the squared error for point estimates of universe scores.

If *I** could be chosen so that the two constants in Equation 19 are small compared to the first two variance components in Equation 11, the expected squared error over *R* for the standardized measurement procedure would be smaller than the expected squared error for the unstandardized procedure, and a biased estimate with a small variance may be preferred to an unbiased estimate with a large variance. Another possibility is to "calibrate" the measurement procedure by estimating the value of α_{I*} and subtracting this value from all observed scores. However, there are serious problems in estimating α_{*} (Cronbach et al., 1972, p. 101); and for most practical applications, standardization cannot be expected to improve the accuracy of point estimates of universe scores.

Standardization is a much more promising approach when universe scores are estimated relative to the expected universe score in the population. If all observations have *I** as the conditions of the *i* facet, the expected value of the observed score over the population is

$$\mu_{I*} = \mu + \alpha_{I*} \qquad [20]$$

and, assuming that μ_{I*} is known, the error in estimating universe deviation scores from observed deviation scores is

$$\delta_{oI*R} = (X_{oI*R} - \mu_{I*}) - (\mu_o - \mu) = \alpha_{oI*} + \alpha_R. \qquad [21]$$

The main effect α_{I*} does not appear in Equation 21 because it is a constant for all observed scores and therefore has no effect on the differences between observed scores and the expected observed score.

The expected value of δ_{oI*R}, over repeated applications of the standardized measurement procedure is given by

$$\underset{R}{\xi}(\delta_{oI*R}) = \alpha_{oI*}. \qquad [22]$$

Therefore, the standardized measurement procedure is also biased in its estimates of universe deviation scores; but the magnitude of the bias, consisting only of the interaction effect, α_{oI*}, is smaller than it is for point estimates of universe scores. For a standardized measurement procedure, the *oI** interaction is a constant for each object of measurement. Therefore, unless α_{oI*} is zero for all objects of measurement, universe deviation scores are systematically overestimated for some objects of measurement and are systematically underestimated for others.

The expected value, over replications, of the squared error in estimating the universe deviation score for object, *o*, is

$$\underset{R}{\xi}(\delta^2_{oI*R}) = (\alpha_{oI*})^2 + \sigma^2(R). \qquad [23]$$

If an *I** with a small *oI** interaction were available, the expected squared error could be reduced even further, but this is usually not practical. The expected value of Equation 23 over the possible choices for *I** is

$$\underset{I*}{\xi}\,\underset{R}{\xi}(\delta^2_{oI*R}) = \sigma^2(oI) + \sigma^2(R), \qquad [24]$$

which is smaller than the expected squared error δ^2_{oIR} for the unstandardized procedure. Therefore, standardization tends to decrease the errors in estimates of deviation scores even if standardization is on randomly chosen conditions of the *i* facet.

The main advantage in standardization is that it can reduce the error variance. Standardization is most useful when observed scores are used to estimate universe deviation scores and the main effect variance for the standardized facet is relatively large. Standardization of the *i* facet eliminates $\sigma^2(I)$ from the error variance for estimates of universe deviation scores but not from the error variance for point estimates.

Systematic Errors

Standardization can be a powerful tool for controlling errors of measurement for universe deviation scores, but there is a price to be paid for this reduction in error variance. If the conditions of the *i* facet are the same for all observations, the effects α_{I^*} and α_{oI^*} are constants over replications of the measurement procedure. For a given object of measurement, therefore, standardization changes some effects from random variables to constants. Components of the error that are constant for all observations on an object of measurement are called *systematic errors* (Cronbach et al., 1972, p. 358). The effect α_{I^*} is a *general systematic error*, since it is a constant over all observations for all objects of measurement. The interaction effect, α_{oI^*}, is a *specific systematic error*, which is a constant for each object of measurement but may vary from one object of measurement to another.

Systematic errors differ from random errors in two ways. First, since their expected value over replications of the standardized measurement procedure is not zero, the systematic errors introduce bias into estimates of the universe scores. The main effect α_{I^*} is the same for all objects of measurement and represents a general bias, which is present in Δ but not in δ. The interaction effect, α_{oI^*}, is a specific bias for each object of measurement, *o*, and it affects both kinds of errors. Since the systematic errors are constants for each object of measurement, they do not tend to "cancel out" over a series of observations; and therefore increasing the number of replications does not decrease the systematic error variance.

Second, the systematic errors are correlated across administrations of the measurement procedure. Since the expected value of Equation 18 over the population is α_{I^*}, the expected value over *I* of the covariance between the errors Δ on two independent administrations of the standardized procedure is given by

$$\underset{I^*}{\xi}\operatorname{cov}(\Delta_{oI^*R}, \Delta_{oI^*R'}) = \underset{I^*o}{\xi}\,\xi(\alpha_{oI^*} + \alpha_R)(\alpha_{oI^*} + \alpha_{R'})$$

$$= \underset{I^*}{\xi}\,\sigma^2(oI^*) = \sigma^2(oI). \quad [25]$$

Similarly, the expected covariance between the errors δ on two independent administrations of the standardized measurement procedure is also given by $\sigma^2(oI)$. Thus, for the standardized procedure, both types of errors of measurement are correlated, and the average correlation depends on the variance of the specific systematic errors (see Lord & Novick, 1968, page 181). Since the systematic errors are correlated across observations, they increase the consistency among observations for each object of measurement and therefore increase reliability while decreasing validity.

The Universe of Allowable Observations and Reliability

In standardizing the *i* facet by requiring that every measurement involve the conditions I^*, a new kind of universe, the *universe of allowable observations*, is introduced. The universe of allowable observations is a subset of the universe of generalization and includes all observations in the universe of generalization that have the appropriate condition for each standardized facet. An instance of the standardized measurement procedure is a randomly sampled observation from the universe of allowable observations, which defines the measurement procedure in the same way that the universe of generalization defines an attribute (both are "extensive" definitions). By constrast, an instance of the unstandardized measurement procedure is a randomly sampled observation from the full universe of generalization.

The question of validity has been examined in terms of how well observations estimate the universe score, or equivalently, how well the measurement procedure satisfies the invariance laws associated with the attribute. As a result of standardization, however, the observations that are actually used to estimate universe scores are drawn from a subuniverse of the universe of generalization. A natural question to ask, then, is how well the observed scores generalize to this universe of allowable observations. This question is equivalent to the question of how well repeated applications of the procedure (i.e., repeated samples of observations from the universe of allowable observations) agree with each other, and this issue is usually treated under the heading of reliability. Therefore, *reliability* is defined in terms of the universe of allowable observations:

> A measurement procedure is reliable to the extent that its observed scores provide dependable estimates of the expected value over the universe of allowable observations.

According to this definition, a measurement procedure is reliable if the observed scores for each object of measurement cluster around the expected observed score for repeated application of the standardized procedure. This definition is consistent with classical test theory if the "true score" is defined as the expected value over the universe of allowable observations.

Reliability is defined as a property of a measurement procedure and does not depend on the definition of the attribute. As noted earlier, validity depends on both the measurement procedure and the attribute being measured. Reliability provides an index of consistency among observed scores, and validity provides justification for an interpretation of observed scores in terms of the universe of generalization defining an attribute.

Random Errors and Reliability Coefficients

Since the attribute is defined in terms of the universe of generalization, measurements of the attribute involve inferences to the universe of generalization rather than to the universe of allowable observations. However, the reliability of the standardized procedure is also of interest; therefore, the dependability of inferences from observed scores to the expected value over the universe of allowable observations, μ_o^*, is often examined. In doing so, the *i* facet is treated as a fixed effect.

If the *i* facet were treated as fixed, the variance components for the *i* facet would be confounded with those for the other facets and would not be estimated independently. If the *i* facet is not included in the analysis of the G study, the *i* facet would be what Cronbach et al. (1972, p. 122) call a "hidden facet." With the *i* facet fixed, the universe score variance is

$$\sigma^2(\mu_o^*) = \sigma^2(o) + \sigma^2(oI^*) + 2cov(o,oI^*). \qquad [26]$$

The expected value of the covariance term in Equation 26 is zero; but for any particular I^*, this covariance may be either positive or negative. Subsequent development and discussion is simplified considerably by taking the expected value of Equation 26 over conditions of the *i* facet:

$$\underset{I^*}{\xi}\,\sigma^2(\mu_o^*) = \sigma^2(o) + \sigma^2(oI). \qquad [27]$$

For generalization over the universe of allowable observations, the expected universe score variance is given by Equation 27 and the error variance is given by $\sigma^2(R)$. For inferences to μ_o^*, the generalizability coefficient is

$$\xi\rho^2(X_{oI*R},\mu_o^*) = \frac{\sigma^2(o) + \sigma^2(oI)}{\sigma^2(o) + \sigma^2(oI) + \sigma^2(R)} . \qquad [28]$$

This coefficient is approximately equal to the expected squared correlation, over I^* and R, between observed scores and the expected value over the universe of allowable observations. This coefficient assumes that generalization is over replications, but not over the i facet, and it indicates the expected consistency of observed scores for a standardized measurement procedure. Therefore, Equation 28 can be interpreted as a reliability coefficient, but it represents an "average" reliability over I^*, rather than the reliability for a specific standardized measurement procedure.

The reliability coefficient in Equation 28 does not estimate the dependability of inferences to μ_o, the expected value over the universe of generalization defining the attribute. A generalizability coefficient for inferences to μ_o is given by the ratio of the universe score variance in Equation 9 to the expected observed score variance for a standardized measurement procedure:

$$\xi\rho^2(X_{oI*R},\mu_o) = \frac{\sigma^2(o)}{\sigma^2(o) + \sigma^2(oI) + \sigma^2(R)} . \qquad [29]$$

Since Equation 29 reflects the expected agreement between observed scores and the value of the attribute, μ_o, it can be interpreted as a validity coefficient. Equation 29 is the expected validity over I^*.

The reliability, as given by Equation 28, is limited by the magnitude of the random errors only. If the i facet is standardized, the interaction effect, α_{oI^*}, is a systematic error and contributes to the numerator of Equation 28 as well as to its denominator. Since variance components are positive, the reliability index in Equation 28 is always greater than or equal to the validity index in Equation 29. For the sampling model, this well-known result from classical test theory (i.e., that reliability is an upper bound for validity) can be interpreted as reflecting the fact that inferences to the universe of allowable observations are generally more dependable than inferences to the more broadly defined universe of generalization.

In evaluating measurement procedures, it will often be necessary to work with partial information, because most universes of generalization have many facets and only a few facets can be systematically investigated in any G study. Since large sample sizes are generally needed for the accurate estimation of variance components in designs with as few as two facets (Smith, 1978), an adequate analysis of the generalizability of a measurement procedure will typically require a number of G studies. Although a G study that does not estimate the variance component for the i facet does not provide enough information to estimate the validity index in Equation 29, it may provide enough information to estimate the reliability index in Equation 28.

Systematic Errors and Validity

The difference between Equation 29, which is interpreted as a validity coefficient, and Equation 28, which is interpreted as a reliability coefficient, is in the role played by $\sigma^2(oI)$. Equation 25 states that $\sigma^2(oI)$, the specific systematic error variance, is the expected covariance of the errors of measurement over repeated observations on the object, o. As $\sigma^2(oI)$ increases, the covariance between the observed scores on two independent administrations of the standardized measurement procedure increases, and thus the reliability increases. However, as $\sigma^2(oI)$ increases, the validity coefficient in Equation 29 decreases. By contrast, both the reliability and validity are decreasing functions of $\sigma^2(R)$, the random error variance.

For a standardized measurement procedure, taking the expected value over repeated measurements implies taking an expected value over R but not over I; and the expected value over the universe of allowable observations, μ_o^*, is analogous to the "true" score of classical test theory. In the limit, as the number of replications approaches infinity, the magnitude of the random errors approaches zero, and the observed score approaches μ_o^*. Therefore, μ_o^* is a parameter for which the standardized measurement procedure provides unbiased estimates. Since the measurement procedure is intended to provide estimates of the universe score, μ_o, the correlation between μ_o and μ_o^* provides an index of the agreement between what the procedure actually estimates without bias and what it is intended to estimate. The expected value over I of the squared correlation between μ_o^* and μ_o is approximately equal to

$$\xi\rho^2(\mu_o^*,\mu_o) = \frac{\sigma^2(o)}{\sigma^2(o) + \sigma^2(oI)} \;. \qquad [30]$$

Equation 30 is approximately equal to the squared correlation between the universe score and an observed score for which the sampling of the universe of allowable observations is sufficiently thorough so that random errors can be ignored; in the limit, as n_r approaches infinity, $\sigma^2(R)$ approaches zero and Equation 29 reduces to Equation 30. Since $\sigma^2(R)$ cannot be negative, Equation 30 provides an upper bound on Equation 29 and therefore provides an upper bound on the validity of the standardized measurement procedure.

Equation 30 can be represented as a validity coefficient corrected for attenuation.

$$\xi\rho^2(\mu_o^*,\mu_o) = \frac{\xi\rho^2(X_{oI*R},\mu_o)}{\xi\rho^2(X_{oI*R},\mu_o^*)} \;, \qquad [31]$$

where the numerator of Equation 31 is the validity coefficient given by Equation 29 and the denominator is the reliability coefficient given by Equation 28. $\xi\rho^2(\mu_o^*, \mu_o)$ represents a disattenuated validity coefficient for the standardized measurement procedure.

The Reliability-Validity Paradox

The inference from the observed score to the universe score can be separated into two parts. The first part is an inference from the observed score to μ_o^*, the expected value over the universe of allowable observations; and the second part is an inference from μ_o^* to μ_o, the expected value over the universe of generalization, or the universe score. The dependability of inferences from observed scores to universe scores can be factored to reflect these two partial inferences:

$$\xi\rho^2(X_{oI*R},\mu_o) = \xi\rho^2(X_{oI*R},\mu_o^*)\;\xi\rho^2(\mu_o^*,\mu_o). \qquad [32]$$

The first factor in Equation 32 is a reliability coefficient and represents the dependability of inferences from observed scores to μ_o^*. The second factor in Equation 32 is a disattenuated validity coefficient and represents the dependability of inferences from μ_o^* to μ_o.

It is clear from Equation 28 that the reliability, which is the first factor in Equation 32, is improved by increasing the impact of the oI interaction variance. This can be accomplished by selecting a single condition for standardization on the i facet. Since the oI interaction is the specific systematic error, this approach to improving reliability has an undesirable consequence; as can be seen from Equation 30, it decreases the disattenuated validity coefficient, which is the second factor in Equation 32 and thereby decreases the overall validity. Therefore, attempts to increase reliability by standardization of the measurement procedure may decrease validity (Lord & Novick, 1968, p. 334).

Convergent Validity

The sampling model analyzes reliability and validity in terms of the evidence for inferences from observed scores. In its emphasis on the justification of inferences, or interpretations, the sampling model is akin to the more general view of validity embodied in construct validity (Cronbach & Meehl, 1955), which is discussed in a later section. At this point, it is convenient to discuss one form of construct validity, namely, convergent validity (Campbell & Fiske, 1959), in terms of the sampling model.

Convergent validity is generally evaluated in terms of the correlations between measurements of an attribute obtained by several different methods. If these correlations are low, the observed scores are seen as being contaminated by method variance, and the validity of any of the methods for measuring the attribute must be doubted. If the correlations are higher, convergent validity is supported. The logic of convergent validity requires that several different methods be available for the measurement of an attribute and therefore assumes that the attribute is a general property not tied to a particular method of observation.

The convergent validity of a measurement procedure can be examined by letting the i facet represent different methods of observation (e.g., objective tests, ratings, observation procedures). A measurement procedure employing a particular method is standardized to that condition of the method facet. With I^* representing a standardized method, the reliability of a measurement procedure can be represented by Equation 28, and its validity can be represented by Equation 29. The validity coefficient in Equation 29 is approximately equal to the expected correlation between scores based on different methods (Kane & Brennan, 1977), and therefore it provides an index of the average "convergent validity" over pairs of methods.

Convergent validity (Campbell & Fiske, 1959) has generally been analyzed in terms of the correlations between specific pairs of methods, rather than the average correlation given by an intraclass correlation coefficient like Equation 29. Boruch and Wolins (1970) have advocated the use of ANOVA to partition observed score variance into components for traits, method, and random error; however, Schmitt, Coyle, and Saari (1977) have criticized the use of the analysis of variance in evaluating convergent validity because it does not provide information about the convergent validity for particular methods. The choice of a specific methodology is not crucial for the sampling model. Although this paper has used analysis-of-variance models to develop the sampling model, correlational methods would be more appropriate whenever the properties of specific methods were of interest.

Convergent validity is basically invariance over methods and depends on the interpretation of the attribute as a general property that is not linked to a specific method of observation. Assuming that I represents a method, the systematic errors α_{oI} are the object-method interactions. For a particular method I^*, α_{oI^*} represents the specific systematic error that results from using method I^*. A large value for $\sigma^2(oI)$ means that method has a serious effect on the results of measurement, and therefore generalization over methods is not appropriate.

The definition of an attribute implies invariance over all facets in the universe of generalization, including those that are standardized. If all of the invariance properties were tested and verified, then the measurement procedure would be valid. If some of the invariance properties were tested and no violations were detected, the validity of the procedure would be partially supported. If even one invariance property were seriously violated, the procedure would be invalid. Convergent validity is essentially invariance over methods.

V. Theory Development

In measuring an attribute, μ_o, for the object, o, the effects α_i and α_{oi} are components of the error. The larger these components are, the more difficult it is to obtain dependable estimates of μ_o; and therefore the effects α_i and α_{oi} are generally viewed as sources of "noise," which need to be controlled. As described in the last section, standardization provides one way of dealing with these errors, but standardization introduces systematic errors. Furthermore, the fact that observations depend on the i facet may be of interest in itself, aside from its effect on inferences to μ_o. Where a functional relationship between observed scores and the i facet can be described by an empirical law, a powerful technique for controlling errors of measurement becomes available. (In this section, the generic term "effect" is used instead of the term "facet," because it is not being assumed here that the i effect is a facet of the universe.)

The errors introduced into measurements of μ_o by the i effect can always be eliminated by shifting attention to a new attribute, μ_{oi}, which involves the same kind of operations that are used to define the attribute μ_o but which has as its objects of measurement the pairs oi instead of the original objects of measurement, o. The universe scores, μ_{oi}, for these new objects of measurement are found by taking the expected value of the observed scores over replications (but not over the i effect, which is now part of the object of measurement rather than being a facet):

$$\mu_{oi} = \underset{R}{\xi}(X_{oiR}) = \mu + \alpha_o + \alpha_{oi} + \alpha_i . \qquad [33]$$

This redefinition of the objects of measurement changes α_i and α_{oi} from being error components to being components of the universe score. If the objects of measurement are defined by both i and o, the difference between the two universe scores, μ_{oi} and $\mu_{oi'}$, involving different conditions of the i effect, is taken as a substantive difference rather than as an error of measurement. Therefore, the i effects and the oi interactions, which account for the difference between μ_{oi} and $\mu_{oi'}$ become part of the universe score.

Measurements of μ_{oi} are more dependable than measurements of μ_o because the interpretation of μ_{oi} is narrower than that of μ_o and thus involves inferences that are less susceptible to errors than those implied by μ_o. While the original attribute, μ_o, characterizes o for all conditions of the i effect, the new attribute, μ_{oi}, characterizes o for a particular condition of the i effect. Therefore, inferences from observed scores to μ_{oi} involve generalization over R but not over i, while inferences to the more general attribute, μ_o, involve generalization over both i and R.

If the observed score X_{oiR} is used to estimate μ_{oi}, the only source of error is the replication facet, and the dependability of inferences from X_{oiR} to μ_{oi} is given by

$$\xi\rho^2(X_{oiR},\mu_{oi}) = \frac{\sigma^2(o) + \sigma^2(oi) + \sigma^2(i)}{\sigma^2(o) + \sigma^2(oi) + \sigma^2(i) + \sigma^2(R)} . \qquad [34]$$

Equation 34 is approximately equal to the expected value of the squared correlation between the observed score X_{oiR} and the universe score μ_{oi}.

When the universe of generalization is restricted to a particular condition of the i effect, μ_{oi} becomes the universe score, and Equation 34, which reflects the dependability of inferences from X_{oiR} to μ_{oi}, is a validity coefficient, with the oi combinations as the objects of measurement and generalization over R. This validity coefficient $_\xi\varrho^2(X_{oiR}, \mu_{oi})$ is never less than the validity coefficient $_\xi\varrho^2(X_{oiR}, \mu_o)$ with objects of measurement, o, and generalization over i and R. Restricting the universe of generalization improves the validity of measurement whenever $\sigma^2(oi)$ or $\sigma^2(i)$ is greater than zero. (Under the same conditions, standardization of the i facet improves reliability but does not necessarily improve validity.)

The increase in validity obtained by restricting the universe of generalization is part of a trade-off by which errors of measurement are reduced but the interpretation of the attribute is narrowed. A high value for Equation 34 indicates that an inference from the observed score X_{oiR} to the universe score μ_{oi} is dependable and therefore provides justification for such inferences. However, if Equation 34 is to be a validity coefficient, inferences must be limited to the universe score μ_{oi} for a specific value of i. The value of Equation 34 does not indicate the dependability of inferences from μ_{oi} to $\mu_{oi'}$, the universe score for the same value of o and a different value of i, or to μ_o, the expected value of μ_{oi} over all values of i. Therefore, a high value for $_\xi\varrho^2(X_{oiR}, \mu_{oi})$ provides support for a relatively limited set of inferences.

The expected squared correlation between the universe score μ_{oi} and the universe score μ_o is approximately equal to

$$_\xi\rho^2(\mu_{oi},\mu_o) = \frac{\sigma^2(o)}{\sigma^2(o) + \sigma^2(oi) + \sigma^2(i)} . \qquad [35]$$

Equation 35 can be derived by substituting "i" for "I" in Equation 14 and setting $\sigma^2(R)$ equal to zero. That is, Equation 35 provides an index of dependability of inferences to μ_o for observed scores based on a single condition of the i facet and an infinite number of replications. Furthermore, by comparing Equations 14, 34, and 35, it is clear that

$$_\xi\rho^2(X_{oiR},\mu_o) = {_\xi\rho^2}(X_{oiR},\mu_{oi})\ {_\xi\rho^2}(\mu_{oi},\mu_o). \qquad [36]$$

Equation 36 partitions the dependability of inferences from X_{oiR} to μ_o into two parts. The first part, $_\xi\varrho^2(X_{oiR}, \mu_{oi})$, represents the dependability of inferences from X_{oiR} to μ_{oi}, the expected value over replications for a particular value of i. The second part, $_\xi\varrho^2(\mu_{oi},\mu_o)$, is the dependability of inferences from μ_{oi} to μ_o. For the investigator who intends to generalize to the universe score, μ_o, therefore, there is no benefit in fixing the condition of the i effect. As a matter of fact, the dependability of inferences to μ_o would be improved by explicitly recognizing the i effect as a facet and by sampling it more thoroughly.

The main benefit derived from restricting the universe of generalization is the increase in validity or the dependability of inferences from observed scores to universe scores. The main disadvantage in restricting the universe of generalization is that it can lead to a large increase in the number of objects of measurement in the population. If there were N_o objects in the original population and N_i conditions of the i effect, there are $N_o\ N_i$ objects in the new population.

The choice between the more narrowly defined attribute, μ_{oi}, and the more broadly defined attribute, μ_o, must be made on pragmatic grounds; as Kaplan (1964, p. 77) has put it, "it is easy to sharpen concepts as much as we like; what is hard is to determine whether this sharpness is worth achieving in a particular way." For example, in analyzing data for a group of persons on several forms of a test, the investigator doing a generalizability study would often assume that the test forms are sampled from a universe of test forms and would generalize over test forms to μ_o. However, the investigator could generalize to the universe scores μ_{oi} for specific test forms, i, and thereby treat the scores on the different test forms as separate variables. Covariance structure analysis (Jöreskog, 1978; Linn

& Werts, 1979) provides one way to analyze the relationships among the different test forms. Covariance structure analysis seeks to estimate the reliability of particular measures by analyzing the variances and covariances for these measures in terms of a specific model for the structure of the measures.

In general, the investigator who views the attribute broadly enough so that the different conditions of the effect are associated with the same attribute will generalize to μ_o, while the investigator who is interested in the relationship among scores for specific conditions of the *i* effect will generalize to μ_{oi}, for the different test forms.

Inferences That Go Beyond the Sampling Model

In some cases, it is possible to characterize how μ_{oi} depends on *i* and *o* by developing an empirical law of the form:

$$\mu_{oi} = f(v_o, w_i), \qquad [37]$$

where *f* represents some function, v_o is a variable that depends on *o* but does not depend on *i*, and w_i is a variable that depends on *i* but not on *o*. An important special case of Equation 37 has the following form:

$$\mu_{oi} = g(\mu_{oi*}, \; w_i - w_{i*}), \qquad [38]$$

where *g* represents some function, and *i** is a particular condition of the *i* facet. A new variable, μ_{oi^*}, is defined in terms of a specific reference condition, *i**, for the *i* effect. This new variable can be substituted for v_o because it is a function of *o* but not of *i*. The fact that measurements of μ_{oi} and of μ_{oi^*} are more dependable than measurements of μ_o facilitates the development of laws of the form given by Equation 38.

If a law like that in Equation 38 can be developed, the limitation inherent in measurements of μ_{oi} can be overcome. With the help of Equation 38, information about μ_{oi^*} for any object provides information about μ_{oi} for all conditions of the *i* effect for which w_i is known. Inferences from observations involving one condition of the *i* effect to what would be expected for another condition of the *i* effect are more difficult to develop than inferences based on invariance laws, but this more complicated approach provides a more detailed analysis of the relationship between *o* and *i*.

The law of thermal expansion, relating length to temperature provides a good example of the kind of law indicated by Equation 38. Since variations in temperature generate errors in measurements of length, the accuracy of measurement can be improved by defining a new quantity, l_{bt}, as the length of a bar, *b*, at temperature *t*. The object of measurement for l_{bt} is a bar-temperature combination, *bt*, instead of a bar. The attribute l_{bt} has a smaller universe of generalization than the attribute l_b, and the direct interpretation of measurements of l_{bt} are restricted to the temperature *t*. However, this restriction is effectively eliminated by the law of thermal expansion, which can be written as

$$l_{bt} = l_{bt*} + k(t - t*)l_{bt*}, \qquad [39]$$

where *t** is some fixed reference temperature. (For convenience, *t** is often taken to be 20°C, a comfortable value for room temperature.) For a fairly wide range of temperature, the coefficient of thermal expansion, *k*, is a constant. Because temperature variations introduce error into measurements of l_b, l_{bt^*} can be measured more dependably than l_b. Since temperatures can be measured very accurately, and since Equation 39 provides a good fit to data over a wide range of temperatures, fixing the temperature for measurements of length does not seriously limit the interpretation of these measurements.

Examples of the model in Equation 37 are provided by latent trait theories (Lord, 1980; Wright & Stone, 1979), which represent the probability that a person answers an item correctly in terms of the ability of the person and one or more item parameters. The ability parameter is an attribute of the person and is assumed to depend on the person but not on the sample of items used to estimate it. The item parameters are attributes of items and are assumed to depend on the items but not on the sample of persons used to estimate them. Once the ability and item parameters for the latent trait model are estimated, an equation like Equation 37 can be used to estimate the performance of each person to each item.

Note that an invariance property is a special case of Equation 38. In particular, if the function, g, is such that μ_{oi} is a constant for all values of i, μ_{oi} is invariant with respect to the i facet. In such cases, there is no loss involved in taking o, instead of oi, as the object of measurement and there is some gain in simplicity. (In practice, it is often convenient to assume that μ_{oi} is invariant with respect to the i facet, even where this assumption is known not to hold exactly.)

The Bridge Analogy

The observations involved in measurement are of interest mainly because they support inferences. These inferences are of two kinds. First, there is an inference from the observation to the universe score. Second, there are inferences from one universe score to other universe scores and to other observations. Using the bridge analogy of Cornfield and Tukey (1956, p. 912), these two inferences can be viewed as two spans of a bridge crossing a river. The first span represents inferences from the observed score to a universe score, and the second span represents inferences from the universe score to the universe scores for other attributes. If the second span is firmly supported by empirical laws, it may be made quite long without weakening the total inference, thus making it profitable to shorten the first span by narrowing the universe of generalization.

Inferences from observed scores to the universe score μ_{oi} for the restricted universe of generalization have a higher validity than inferences to μ_o. Therefore, restricting the universe strengthens the first span. A well-confirmed law of the type given in Equation 38 provides a strong second span by justifying inferences from one universe score to another. In such cases, restricting the universe of generalization does not limit the generality of the inferences; the second span is simply bearing more responsibility for the total inferences.

The Import of Measurements

In discussing inferences that go beyond the universe of generalization, it is useful to define a third property of measurement in addition to reliability and validity. This third property, the *import* of measurement, is associated with the total significance of what can be inferred from the measurement. Hempel (1952) introduced import in terms of an example:

> . . . we might define the hage of a person as the product of his height in millimeters and his age in years. This definition is operationally adequate and the term "hage" thus introduced would have relatively high precision and uniformity of usage; but it lacks theoretical import, for we have no general laws connecting the hage of a person with other characteristics. (p. 46)

Import is a qualitative concept, which emphasizes the scope and significance of the inferences that can be drawn from a measurement.

The invariance laws that justify the interpretation of observations as measurements provide a core of import to all measurements. These laws support inferences from the observed scores to the universe score and, also, to all other observed scores in the universe of generalization. If the attribute is not involved in any other empirical laws, the invariance laws define the total import of the measurement. The extent to which the invariance laws lend import to measurements depends in part on the generality of these laws. If attention is restricted to the invariance properties implied by their definitions, the attribute μ_o has greater import than the attribute μ_{oi}. As such, measurements of μ_{oi} do not justify inferences to other conditions of the *i* facet. Measurement of μ_o, on the other hand, involves generalization over all conditions of the *i* facet.

In practice, the development of empirical laws can lead to simultaneous increases in both the validity and the import of measurements. This is accomplished by partitioning the universe of generalization into a number of more narrowly defined subuniverses, while connecting the universe scores for these subuniverses through empirical laws. Physics has used this strategy very effectively. Measurement of such basic attributes as length have been gradually refined by restricting various facets in their universe of generalization. At the same time, the import of length measurements has been increased by theories like Euclidian geometry and classical mechanics.

The Role of Theories

Attributes that play a central role in fundamental theories have greater import than attributes which are involved in one or two isolated empirical laws, or in no laws at all. The empirical laws that determine the content of a theory and provide confirmation for the theory add to the import of all of the attributes involved in the theory. Theories extend the range of inferences that may be drawn from measurements and therefore increase their import. However, according to the sampling model, the validation of measurements that are interpreted as dispositions does not depend on theory.

Measurements of a disposition are valid to the extent that they provide accurate estimates of universe scores. The existence of laws or theories involving a dispositional attribute has no direct bearing on the validity of measurements of the attribute. For example, within atomic theory, the law of thermal expansion, Equation 39, can be at least partially explained in terms of the motion of molecules. However, the theory is not necessary in order to interpret the coefficient of thermal expansion, *k*, which is simply the rate at which length changes as a function of temperature. The theory provides a causal explanation of thermal expansion but is not needed for the interpretation of *k* as a dispositional attribute and is therefore not used in validating measurements of *k*.

This point of view is generally consistent with the interpretation of measurements in science. In his analysis of measurement in physics, Campbell (1921, p. 134) concluded that "measurement is essential to the discovery of laws" but he did not use the laws to evaluate measurement procedures. Similarly, Suppes and Zinnes (1963) and, more recently, Krantz, Luce, Suppes, and Tversky (1971) did not find it necessary to consider most of the laws involving attributes in their detailed analyses of measurement.

Although they do not have a direct role in validating measurements for dispositional attributes, theories have a significant indirect influence on the validity of such measurements. They make it feasible to restrict the universe of generalization for attributes are therefore to increase the validity of measurements without decreasing their import. These more narrowly defined attributes depend on the theories for much of their import, while the magnitude of the errors of measurement are reduced because fewer invariance properties are required.

The Tradeoff Between Validity and Import

If import is ignored, it is easy to generate measurements with a high degree of validity by defining the universe of generalization narrowly enough so that the inferences to the universe scores involve generalization over few facets. Such inferences are likely to be very dependable. In the limiting case, where observations are interpreted simply as observations, there is no inference and, therefore, no chance of an invalid inference.

However, the issue of import cannot be ignored, and trade-offs between validity and import must be made; Kaplan's (1964, p. 77) question about the appropriate degree of sharpness for concepts must be addressed. The researcher who interprets observations narrowly draws more accurate inferences but also says less about the world than the researcher who interprets observations broadly. The choice between narrow but dependable interpretations and broader, but less dependable, interpretations is a choice of strategy. The continuum of available options has strict operationalism at one end and construct validity at the other end.

Strict operationalism tries to define attributes narrowly enough to insure that the validity of the interpretations is essentially perfect (Bechtoldt, 1959). The strict operationalist is unwilling to give any hostages to the future in the form of assumed invariance properties that might turn out to be only approximations. Strict operationalism is the strategy of pure empiricism, and theory plays essentially no role. As described by Cronbach and Meehl (1955), construct validity defines an attribute in terms of all of the relationships in which it appears. From the standpoint of construct validity, the definition of an attribute entails various laws; and in order for a measurement procedure to be valid, all of these laws must be satisfied.

Construct Validity

According to the American Psychological Association Standards (American Psychological Association, 1974), a psychological construct is "a theoretical idea developed to explain and to organize some aspects of existing knowledge," and construct validity occurs "when one evaluates a test or other set of operations in light of the specified construct" (p. 29). Cronbach and Meehl (1954) summarized the logic of construct validity by saying,

> Construct validation takes place when an investigator believes that his [or her] instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim. (p. 255)

A dispositional attribute is defined by its universe of generalization, and the laws involved in giving meaning to such constructs are the invariance properties. The sampling model assumes that the interpretation of a measurement generates testable hypotheses—the invariance properties, which can be used to test the validity of the interpretation.

The sampling model is generally consistent with construct validity. Like construct validity, it suggests that validity is more accurately represented as a series of upper bounds than as a single number. Since most attributes assume invariance over a number of facets, it will usually require a series of studies to establish invariance over all of the potentially important facets. If one or more of the required invariance properties does not hold, there are two general options available (see Cronbach & Meehl, 1955, p. 260-264). First, the measurement procedure can be modified, possibly by sampling more thoroughly or by standardizing some facets. Second, the attribute's definition can be changed by restricting the universe of generalization.

Although the sampling model is consistent with construct validity in most ways, it is less general in the interpretations to which it can be applied. The sampling model emphasizes the interpretation of attributes as dispositions, which are defined in terms of universes of observations, while construct validity applies to a variety of constructs ranging from theoretical attributes embedded in extensive networks to basic dispositions. Theoretical constructs derive their meaning from the axioms of a theory (Braithwaite, 1953) and are said to be defined implicitly by the network of laws implied by the axioms of the theory. Measurements of an implicitly defined construct are validated by verifying the theory. Because Cronbach and Meehl (1955) tend to emphasize the role of theory in providing an implicit definition for theoretical constructs, they suggest that all of the laws involving an attribute should be used in validating measurements of the attribute.

Because construct validity may apply to any kind of interpretation, it is more general than the sampling model. However, the dispositional interpretations emphasized by the sampling model are particularly important in science because they are especially relevant to the central activity of science—theory testing. In order to avoid circularity, the measurements that are used to test a theory must not depend on the theory for their validity. Therefore, in testing theories, dispositional interpretations are typically emphasized in place of more theory-laden interpretations.

The observations used to measure a theoretical attribute can generally be interpreted on at least two levels. First, the observations can be interpreted as samples from the universe of generalization defining a disposition. The sampling model suggests methods that are appropriate for evaluating the validity of a dispositional interpretation. Second, the observations can be interpreted as reflecting a theoretical construct which is implicitly defined by a theory. Although Cronbach (1980b) has suggested that a distinction be drawn between laws involved in the "core meaning" of a construct and other hypotheses that "we are ready to abandon," there are no general criteria for deciding which of the laws in which an attribute appears are most relevant to its construct validity.

Use of the Full Network—a Practical Consideration

In those cases where extensive theory exists, the effort required to test all of the implications of the theory is likely to be prohibitive for any investigator. And since theories in the behavioral sciences are often rather loose, the evidence for validity derived from the theory is likely to be ambiguous at best and inconsistent at worst. By assuming that the theory as a whole implicitly defines constructs, Cronbach and Meehl (1955) present an extremely formidable task to the investigator seeking to establish construct validity, and it would be useful to establish priorities for which laws or theories should be included in validity studies.

The sampling model's distinction between validity and import provides a basis for setting priorities. Validity is defined in terms of invariance over a universe of generalization and is associated with the "meaning" of an attribute in the narrow sense of denotation. Import is defined in terms of all of the relationships in which the attribute is involved and is associated with "meaning" in the broader sense of connotation. The sampling model emphasizes the interpretation of attributes as dispositions and equates validation with the testing of invariance laws; the theory as a whole defines the import of measurements and therefore has a strong, but indirect, effect on validation.

By checking the invariance laws, the investigator establishes that the measurements support inferences of some generality and are not simply reports of isolated incidents. The invariance laws provide the "core meaning," or denotation, for basic attributes. The core meaning of derived attributes includes their defining laws (e.g., the law of thermal expansion for the coefficient of thermal expansion) in addition to the invariance laws. The validation of theoretical attributes, which are implicitly defined by a theory, involves the confirmation of the theory as a whole. However, even for the most theoretical of constructs, the sampling model suggests a reasonable place to begin the study of validity.

VI. The Adequacy of Sampling Models for Validity

The sampling model is basically quite simple. A dispositional attribute is defined in terms of a universe of generalization. The "true" value of the attribute is the universe score, equal to the expected value over the universe of generalization. Measurement involves inferences from observations to the expected value over the universe defining an attribute, and a measurement procedure is valid to the extent that its inferences about universe scores are accurate. For these inferences to be justified, certain invariance laws must hold, at least approximately, and these laws should be verified empirically.

Although the application of the model becomes more complicated when standardization and theory development are considered, the basic premises remain quite simple. The sampling model makes no assumptions about underlying structures or processes; it neither affirms nor denies the existence of such theoretical constructs. It makes no assumptions about the distribution of observed scores, the distribution of universe scores, or the relationships between different kinds of scores. The model does not dictate what kinds of conditions can be defined as facets. The only restriction put on the universe of generalization is that it be sufficiently well defined that it is possible to test the invariance laws associated with the universe.

Objections to the Sampling Model

A number of authors (Gillmore, 1979; Loevinger, 1965; Rozeboom, 1966) have objected to sampling models by pointing out that measurements do not generally consist of random samples from a clearly defined universe of generalization. Within the sampling model, the sequence of inferences involved in going from observed scores to universe scores can be analyzed explicitly, thus making it unnecessary to assume that measurement involves random sampling from the universe of generalization. However, the model does require that universes be defined well enough to allow for the testing of the invariance laws.

Some progress has been made in developing more precise definitions for universes. For example, Anderson (1972), Bormuth (1970), and Popham (1978) have proposed methods for the specification of the item facet for achievement tests, and Fiske (1977) and Wiggins (1973) have discussed the definitions of a variety of facets. However, there is still considerable ambiguity in the universe definitions for scientific attributes. In discussing the semantics of natural languages, Clark and Clark (1977, p. 412) have pointed out that "the boundaries for most categories are fuzzy. There is no clear boundary between trees and bushes; one category shades off into the other." Similarly, Nagel (1971, p. 30) talks about the "penumbra of vagueness" of all scientific terms. The sampling assumptions required for the interpretation of measurement are typically not satisfied exactly; like all models, the sampling model is only an approximation.

In light of the substantial difficulties in drawing random samples from the universe of generalization, the verification of the invariance laws presents a serious problem. The invariance laws are general empirical laws, and their testing involves inductive inferences similar to those involved in the testing of any empirical laws. In practice, a measurement procedure is evaluated in a series of studies in which its invariance laws are examined.

Invariance Laws and Inductive Inference

Even a cursory review of the issues involved in inductive inference would go far beyond the scope of this paper, but some general remarks on this topic may put the problem of testing the invariance laws into perspective. According to Popper (1965, 1968), laws are conjectures which are subject to possible refutation but not to confirmation. A general law applies to many observations, and any of these observations could be used to test the law. A deterministic law that fails a single test or a statistical law that fails a large proportion of its tests is refuted. A law which is subjected to a wide variety of empirical tests without being refuted is supported by these tests.

There is a clear lack of symmetry in Popper's views; a law can be decisively refuted, but even a large number of studies cannot definitely confirm the law. The more challenges of various kinds that a law has been exposed to without being refuted, the more strongly it is considered to be supported, but the law is never completely confirmed. In discussing validation, Cronbach (1980a) has made a similar point:

> The job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it. (p. 103)

In a sense, Popper replaces the concept of confirmation by the concept of degree of confidence. A law is subjected to empirical challenges, and confidence in the law increases whenever it meets a challenge (Lakatos, 1970).

According to the sampling model, the definition of an attribute involves a universe of generalization. The claim that a measurement procedure generates valid measurements of the attribute is equivalent to the conjecture that the observed scores are invariant with respect to sampling from this universe. This conjecture can be decomposed into a number of specific invariance laws, each of which applies to a specific facet. A successful test of any one of the invariance laws, based on a random sample of conditions from the facet, provides direct support for the specific invariance law and some support for the cluster of invariance laws associated with the attribute. As more and more invariance properties are investigated without encountering refutations, the degree of confidence in the validity of the measurements increases.

The strength of the evidence for validity provided by invariance over a particular facet will depend on the facet studied. If there is some reason to suspect that the facet may have a large effect on observed scores, evidence for invariance over that facet answers a serious challenge and therefore provides strong support. Note that G studies, which sample from only part of a facet, do not test invariance over the full facet. They do provide evidence for invariance over the subuniverse sampled. Because they answer one possible challenge to the invariance law, they provide some evidence for validity, but they are not very effective in establishing the limits of the universe.

Since it is usually not possible to estimate variance components for more than a few facets without having very large sample sizes, the validation of most measurement procedures will require a series of studies in which the invariance properties are systematically investigated. Those facets that are expected to introduce the greatest error variance should be investigated first. Subsequently, other facets can be investigated. The resulting sequence of upper bounds on validity is perhaps less satisfactory than a point estimate of validity, but it is probably more realistic to consider any coefficient resulting from a G study as an upper bound than as an unbiased point estimate of validity.

The Procrustes Effect in Defining Universes of Generalization

Throughout most of this paper, it has been tacitly assumed that the conditions defining a facet are given and that the task is to check the appropriate invariance laws. For purposes of exposition, this assumption has been convenient, but in practice the situation is never quite this simple. Although

Popper (1965, 1968) has emphasized the testing of laws, the invariance requirement may also be addressed by defining facets so that the observations are invariant over the facet.

Toulmin (1953) described scientific laws as rules of inference, rather than as inductive generalizations. The question to be asked about such rules of inference is not whether or not they are true, but how widely they apply. Toulmin's analysis is quite different from Popper's; but for the purposes of this paper, these two views are complementary. Toulmin's (1953) treatment of scientific laws as rules of inference fits the sampling model especially well. The whole purpose of the invariance properties is, in fact, to justify inferences from observed scores to universe scores.

As applied to the invariance laws, the task of determining how widely the law applies is closely connected with the task of clarifying the boundaries of the universe of generalization. Toulmin makes the point that the range of applicability of a law is generally not known when the law is first proposed, and one of the aims of empirical research is to determine how widely the law applies. The aim of empirical investigations of the invariance laws is to see how widely observed scores can be generalized.

In most cases, decisions about whether or not to generalize over a class of conditions will depend, in part, on whether observations are invariant over the conditions. If the observations are not invariant over a facet, including the facet in the universe of generalization would decrease the validity of measurements of the attribute; therefore, the facet is not likely to be included in the definition of the universe. However, if the observations are invariant over the facet, including the facet in the universe would not decrease validity and would increase the usefulness of the measurements. The universe of generalization is designed so that the invariance laws will hold. Lumsden (1976, p. 271) makes a similar suggestion.

Random Sampling from the Universe of Allowable Observations

The purpose of G studies is to provide data that can be used in the design of effective measurement procedures. In particular, an important goal in designing a measurement procedure is to reduce the number of facets that must be randomly sampled in obtaining an observed score. There are three ways to do this. First, if G studies show that all of the variance components (for the main effect and interactions) for a facet are zero, there is no need to be concerned about how this facet is sampled. Second, facets that have been standardized are not sampled in estimating universe scores, although the magnitude of the systematic errors for these facets must be estimated in G studies. Third, if the universe of generalization is restricted in connection with the development of theory, thus defining a new attribute, measurements of the new attribute will not involve sampling of the facet.

All of these modifications of the measurement procedure tend to decrease the number of facets that are randomly sampled for each observed score. The only facets that need to be sampled randomly are those for which interactions with the object of measurements are fairly large and apparently random. Efforts to obtain random samples can be concentrated on these facets; and if the number of such facets can be decreased, the difficulty in taking random samples from the universe of allowable observations is reduced.

The Role of Theory in Shaping the Universe of Generalization

The universe of generalization is, to a large extent, shaped by theory. This occurs in at least three ways. First, as discussed earlier, those effects that are explicitly involved in the theory are treated as fixed conditions in the universe of generalization rather than as facets.

Second, theory often suggests the kinds of observations that should be considered equivalent in the sense of belonging to the same universe of generalization. For example, a cognitive psychologist might put the responses to a certain question into the same universe of generalization, whether the question was written, spoken, or projected on a screen and whether the responses were made by speaking, writing, pressing a button, or giving a hand signal. Since the observations in this universe have few surface features in common, it is the theoretical assumptions about cognitive processes that supplies impetus for putting them into the same universe of generalization. One advantage of designing attributes to fit a theory is that the import of the theory is built into the attribute.

Third, instead of defining an attribute in terms of the mean over the universe of generalization or in terms of a particular condition of a facet, it is often useful to define the attribute in terms of an "ideal" limiting condition which never occurs in practice. This avoids having the properties of the measurements determined by any particular condition of observation. The interpretation of human abilities in terms of the best performance, rather than typical performance, involves this kind of ideal condition. Since the best performance can occur only if the person has adequate time to finish the task, abilities are associated with neither a fixed time limit nor with the mean over a time-limit facet. The assumption that the time limit does not affect performance substantially could be investigated by conducting an experiment in which the time limit is extended; if performance improves significantly, the time limit is too short.

The Steady State Requirement

Cronbach et al. (1972, p. 364) have stated that because generalizability theory "treats conditions within a facet as unordered it will not deal adequately with the stability of scores that are subject to trends . . ." and that "the concept of universe score is of dubious value if the universe stretches over a period when the person's status is changing regularly and appreciably." According to the sampling model, to generalize over a facet is to treat the variability of observed scores due to the sampling of the facet as error. If conditions of some kind are considered a facet, the attribute is the expected value over all conditions of the facet and is not associated with any particular condition. For a relationship to exist between observed scores and the conditions of a facet, each observed score must be associated with a particular condition of the facet, and this is not consistent with the interpretation of the facet as a source of error. Therefore, an effect cannot be interpreted as a facet and at the same time as an attribute to be systematically investigated.

The problems associated with trends can be eliminated as soon as the trend is detected; this is accomplished by restricting the universe of generalization for each observation to a fixed condition of the facet involved and by treating the trend as an empirical law. Undetected trends will tend to cause the variance components for the facet to be large, and therefore the examination of variance components can facilitate the detection of trends.

The Analysis-of-Variance as a Tool

Although the analysis of variance is a useful tool for the sampling model, the formal statistical models defining variance components should not be allowed to obscure the fundamental concerns embodied in the invariance properties. As Cronbach (1976) has observed, the technical apparatus of generalizability theory is less important than the questions suggested by the theory. The sampling model provides a framework for considering the issues that arise naturally in the interpretation of measurements. The three types of issues that have been identified are those associated with reliability, validity, and import. For convenience, most of the discussion of these issues has been in terms of variance components, but the same points could have been made using other terminology. In fact, where there is interest in the relationship among the observed scores for particular conditions of a facet, correlation coefficients might be preferred.

In some cases, it may be necessary to employ controlled experiments in investigating the effects of a facet. For example, if an attribute is assumed to be unaffected by short periods of "coaching," as is often the case for aptitudes, it might be necessary to conduct controlled studies to check this assumption and to analyze the resulting data in a way that is consistent with the design of the study. The sampling model requires that the invariance properties implicit in the definition of an attribute be investigated; it does not specify how these invariance properties should be investigated.

Concluding Comments

A dispositional attribute is defined in terms of a universe of generalization, and the universe score for the attribute is the expected value over this universe. Therefore, the universe score is a parameter defined over the universe of generalization, rather than a hypothetical entity. Although the sampling model is basically quite simple, its application can be very difficult for two reasons. First, most measurement procedures are not designed in terms of random sampling from the universe of generalization, and therefore inferences from observed scores to the universe scores may be quite complicated.

The second and more fundamental difficulty is encountered in defining the universe of generalization. As noted earlier, some progress has been made in the methodology for defining universes, but it is still true that the universes associated with most attributes are quite fuzzy. The difficulties involved in defining universes, and subsequently sampling from these universes, arise for measurement procedures in all of the sciences. Violations of sampling assumptions introduce vagueness into what would otherwise be a precise statistical model, but they do not preclude effective use of the sampling model. Rather, the sampling model highlights the weakness of some inferences that are routinely made in interpreting measurements and should therefore encourage research aimed at defining universes more precisely.

Although the sampling model makes few assumptions, it provides an analysis of many issues associated with the dependability of measurement. The sampling model makes it possible to give validity a straightforward interpretation and to draw a clear distinction between reliability and validity. The model provides the basis for a detailed analysis of standardization and of the resulting systematic errors. The conclusions that reliability is an upper bound on validity and that some means of improving reliability may cause validity to decrease are easily derived from the model. Furthermore, the model suggests an explicit mechanism for relating the refinement of measurement procedures to the development of laws and theories.

References

American Psychological Association. *Standards for educational and psychological tests.* Washington DC: Author, 1974.

Anderson, R. C. How to construct achievement tests to assess comprehension. *Review of Education Research,* 1972, *42,* 145-170.

Bechtoldt, M. P. Construct validity: A critique. *American Psychologist,* 1959, *14,* 619-629.

Bormuth, J. K. *On the theory of achievement test items.* Chicago: University of Chicago Press, 1970.

Boruch, R. F., & Wolins, L. A procedure for estimation of trait, method, and error variance attributable to a measure. *Educational and Psychological Measurement,* 1970, *30,* 547-574.

Braithwaite, R. B. *Scientific explanation.* Cambridge, England: Cambridge University Press, 1953.

Brennan, R. L. *Generalizability analysis: Principles and procedures* (ACT Technical Bulletin No. 26). Iowa City IA: American College Testing Program, 1977.

Brennan, R. L. *Elements of generalizability theory* (ACT Monograph 16). Iowa City IA: American College Testing Program, in press.

Bridgman, P. W. *The logic of modern physics.* New York: Macmillan, 1927.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, *56*, 81-105.

Campbell, N. R. *What is science?* London: Methuen, 1921.

Campbell, N. R. *Foundations of science.* New York: Dover Publications, 1957. (Originally published as *Physics: The elements.* Cambridge, England: Cambridge University Press, 1920)

Cardinet, J., Tourneur, Y., & Allal, L. The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, 1976, *13*, 119-134.

Carnap, R. Testability and meaning. In H. Feigl & M. Brodbeck (Eds.), *Readings in the philosophy of science.* New York: Appleton-Century-Crofts, 1953.

Carnap, R. *The philosophy of science* (M. Gardner, Ed.). New York: Basic Books, 1966.

Clark, H. H., & Clark, E. V. *Psychology and language.* New York: Harcourt Brace Jovanovich, 1977.

Corben, H., & Stehl, P. *Classical mechanics* (2nd ed.). New York: John Wiley & Sons, 1960.

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 1956, *27*, 907-949.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington DC: American Council on Education, 1971.

Cronbach, L. J. On the design of educational measures. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement.* New York: John Wiley & Sons, 1976.

Cronbach, L. J. Validity on parole: How can we go straight? *New Directions for Testing and Measurement*, 1980, *5*, 99-108. (a)

Cronbach, L. J. Discussion. In *Proceedings of conference on construct validity in psychological measurement.* Princeton NJ: Educational Testing Service, 1980. (b)

Cronbach, L. J., & Gleser, G. C. The signal/noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement*, 1964, *24*, 467-480.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, *52*, 281-302.

Ebel, R. Must all tests be valid? *American Psychologist*, 1961, *16*, 640-647.

Ebel, R. L. Explorations in reliability theory. *Contemporary Psychology*, 1974, *19*, 81-83.

Ellis, B. *Basic concepts of measurement.* Cambridge, England: Cambridge University Press, 1968.

Ennis, R. H. Operational definitions. In H. S. Broudy, R. H. Ennis, & L. I. Krimerman (Eds.), *Philosophy of educational research.* New York: John Wiley & Sons, 1973.

Fiske, D. W. *Measuring the concepts of personality.* Chicago: Aldine, 1971.

Frank, P. Philosophical interpretations and misinterpretations of the theory of relativity. In H. Feigl & M. Brodbeck (Eds.), *Readings in the philosophy of science.* New York: Appleton-Century-Crofts, 1953.

Gillmore, G. M. *An introduction to generalizability theory as a contributor to evaluation research* (EAC Report No. 79-14). Seattle: University of Washington, Educational Assessment Center, 1979.

Hempel, C. G. *Fundamentals of concept formation in empirical science.* Chicago: The University of Chicago Press, 1952.

Hempel, C. G. Operationism, observation, and theoretical terms. In A. Danto & S. Morgenbesser (Eds.), *Philosophy of science.* New York: New American Library, 1960.

Hempel, C. G. *Aspects of scientific explanation and other essays in the philosophy of science.* Glencoe IL: Free Press, 1965.

Jöreskog, K. G. Structural analysis of covariance and correlation matrices. *Psychometrika*, 1978, *43*, 443-477.

Kaiser, H. J., & Michael, W. B. Domain validity and generalizability. *Educational and Psychological Measurement*, 1975, *35*, 31-35.

Kane, M. T., & Brennan, R. L. The generalizability of class means. *Review of Educational Research*, 1977, *47*, 267-292.

Kaplan, A. *The conduct of inquiry.* San Francisco: Chandler, 1964.

Krantz, D., Luce, R. D., Suppes, P., & Tversky, A. *Foundations of measurement* (Vol. 1). New York: Academic Press, 1971.

Kuhn, T. S. *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press, 1970.

Lakatos, I. Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge.* London: Cambridge University Press, 1970.

Lindquist, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

Linn, R., & Werts, C. Covariance structures and their analysis. *New Directions for Testing and Measurement,* 1979, *4,* 53-73.

Loevinger, L. Person and population as psychometric concepts. *Psychological Review,* 1965, *72,* 143-155.

Lord, F. M. *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum, 1980.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading MA: Addison-Wesley, 1968.

Lumsden, J. Test theory. *Annual Review of Psychology,* 1976, *27,* 254-280.

McDonald, R. P. Generalizability in factorable domains: "Domain validity and generalizability." *Educational and Psychological Measurement,* 1978, *38,* 75-79.

Meehl, P. E. On the circularity of the law of effect. *Psychological Bulletin,* 1950, *47,* 52-75.

Nagel, E. Theory and observation. In E. Nagel, S. Bromberger, & A. Gumbaum, *Observation and theory in science.* Baltimore: The Johns Hopkins Press, 1971.

Nunnally, J. C. *Psychometric theory.* New York: McGraw Hill, 1967.

Physical Science Study Committee. *College physics* (U. Haber-Schain, Ed.). Boston: Heath, 1968.

Popham, W. J. *Criterion-referenced measurement.* Englewood Cliffs NJ: Prentice-Hall, 1978.

Popper, K. R. *Conjecture and refutation: The growth of scientific knowledge.* New York: Harper & Row, 1965.

Popper, K. R. *The logic of scientific discovery.* New York: Harper & Row, 1968.

Rozeboom, W. W. *Foundations of the theory of prediction.* Homewood IL: Dorsey Press, 1966.

Schmitt, N., Coyle, B. W., & Saari, B. B. A review and critique of analyses of multitrait-multimethod matrices. *Multivariate Behavioral Research,* 1977, *12,* 447-478.

Smith, P. L. Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Statistics,* 1978, *3,* 319-346.

Stallings, W. M., & Gillmore, G. M. A note on "accuracy" and "precision." *Journal of Educational Measurement,* 1971, *8,* 127-129.

Suppes, P. The structure of theories and the analysis of data. In W. F. Suppe (Ed.), *The structure of scientific theories.* Urbana IL: The University of Illinois Press, 1974.

Suppes, P., & Zinnes, J. L. Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1). New York: John Wiley & Sons, 1963.

Toulmin, S. *The philosophy of science.* London: Hutchinson's Universal Library, 1953.

Tryon, R. C. Reliability and behavior domain validity; reformulation and historical critique. *Psychological Bulletin,* 1957, *54,* 229-249.

Wiggins, J. S. *Personality and prediction: Principles of personality assessment.* Reading MA: Addison-Wesley, 1973.

Winer, D. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Wright, B. D., & Stone, M. H. *Best test design.* Chicago: Mesa Press, 1979.

Author's Address

Send requests for reprints or further information to Michael T. Kane, American College Testing Program, P.O. Box 168, Iowa City IA 52243.