INTERNATIONAL JOURNAL OF TESTING, *1*(3&4), 283–303

# The Geometry of Probability, Statistics, and Test Theory

Donald W. Zimmerman
*Department of Psychology*
*Carleton University*

Bruno D. Zumbo
*Department of ECPS*
*University of British Columbia*

The model of tests and measurements outlined in this article identifies test scores with Hilbert space vectors and true and error components of scores with linear operators. The collection of all observed scores associated with a test procedure is represented by the function space $L_2(\Omega,A,P)$; the collection of all true scores is the Hilbert subspace $L_2(\Omega,B,P)$, where B is a $\sigma$-algebra contained in A; and the collection of all error scores is the orthogonal complement of the subspace of true scores. This geometric formalism simplifies derivations in test theory and brings to light relations among concepts in probability, statistics, and measurement that are not otherwise apparent. Test reliability, test validity, error of measurement, parallel tests, and other familiar concepts can be studied in this framework, making their mathematical properties and their interrelations with one another more obvious.

This article formulates classical test theory as an abstract mathematical model, using concepts in measure theory, probability theory, and functional analysis. It identifies test scores with vectors in an infinite dimensional vector space, or Hilbert space, and true and error components of scores with linear operators. Correlations among components of scores are inner products between vectors, and various test–theory concepts arise naturally from properties of operators and inner products. This geometric point of view brings to light relations among elementary concepts in test theory, including reliability, validity, and parallel tests, which have not been evident in the classical theory of mental tests.

Requests for reprints should be sent to Donald W. Zimmerman, 1978 134A Street, Surrey, B.C., Canada V4A 6B6. E-mail: zimmerma@look.ca

At first glance, the definitions of true score and error score presented in this article may appear unnecessarily abstract and complicated. Later, however, it becomes apparent that they accomplish precisely what is intended and simplify the theory as a whole. The model in this article is related to the theory of conditional expectation (Doob, 1953; Kolmogorov, 1933; Steyer, 1988, 1989; Steyer & Schmitt, 1990; Zimmerman, 1975), as well as to other "geometric" models in probability and statistics that are based on the same formalism (see, for example, Loève, 1963; Rényi, 1970).

The theories of tests and measurements presented by Cronbach, Rajaratnam, and Gleser (1963), Guttman (1945), Lord and Novick (1968), Novick (1966), Rozeboom (1966), and others, explicitly defined observed scores, true scores, and error scores as random variables having designated properties. These formulations improved on the less systematic approach that had prevailed earlier in the century, deriving from the pioneering work of Spearman and Yule, which was summarized by Gulliksen (1950). For most purposes, the identification of test scores with random variables is all that is needed to develop the theory and to make available the mathematics of probability and statistics. However, the distinctive character of test theory and its relations with other mathematical models becomes more evident when it is incorporated into an abstract framework underlying all these disciplines. This approach is undertaken in this article.

A collection of random variables with finite variance, or square-integrable random variables, defined on a probability space, is an example of the Hilbert space customarily denoted by $L_2(\Omega,A,P)$ in functional analysis and probability theory. The test–theory concepts of true score and error score can be associated in a natural way with projection operators in this Hilbert space. Once this identification is made, metric concepts of distance, length, angle, and orthogonality have immediate applications to test theory.

One purpose of this article is to show that the familiar equations of classical test theory can be examined from several increasingly abstract points of view: (a) as relations among variances, covariances, and correlations of components of scores, which have been studied for many decades; (b) as properties of conditional expectations of random variables defined on probability spaces (Steyer, 1988, 1989; Zimmerman, 1975); and (c) as properties of linear operators defined on function spaces, the elements of which are random variables representing test scores. This article emphasizes advantages of the third approach. The higher the level of abstraction, we shall see, the more comprehensive is the unification of diverse points of view.

This approach yields insight into the meaning and significance of test–theory concepts and brings to light some relations that have not been apparent in traditional models. Another advantage is that mathematical proofs are simplified. Some derivations based on variances and covariances, which are algebraically tedious in the classical theory, can be bypassed by simply observing that a theorem is an immediate consequence of an already-known result in linear algebra or functional analysis. For example, some identities relating score components that have been painstakingly derived in test theory (Gulliksen, 1950; Lord and Novick, 1968) are nothing more than trigonometric identities in another guise (see also Jackson, 1924).

The association of statistical concepts with properties of vectors and linear operators is commonplace in many branches of probability and statistics, including the theory of stochastic processes. Mathematical models based on linear operators also have been prominent in quantum mechanics. The mathematical formalism of quantum theory has been highly successful in organizing and explaining known results, as well as predicting new experimental findings in physics. When first introduced into physics, Hilbert space concepts unified what had previously appeared to be two separate and distinct theories—Heisenberg's matrix mechanics and Schrödinger's wave mechanics. These theories turned out to be mathematically equivalent when reformulated in a Hilbert space setting by Von Neumann, Dirac, and others (see, for example, Cohen, 1989; Messiah, 1959).

Advantages of formulating test theory as an operator model are somewhat similar. Sets of axioms that at first glance appear to be quite different turn out to be equivalent. As already mentioned, the general framework facilitates derivation of many widely known algebraic identities relating components of test scores in very few steps. The operator model highlights a similarity of true scores and error scores that is not conspicuous in the usual definition of an error score as the difference between an observed score and a true score (see definitions 1 and 2 later). In addition to its advantages in organizing classical test theory, the operator formalism also provides insight into extensions and modifications of the classical theory, including generalizability theory, item response theory, and various models based on factor analysis.

The operator formalism, we shall see, also throws light on the meaning of test reliability and the role of parallel measurements in reliability theory. Furthermore, the association of a linear operator with an error score makes clear the significance of correlated errors of measurement (Rozeboom, 1966; Zimmerman & Williams, 1979, 1980). A definition of test validity, which arises naturally from these concepts, embodies the meaning of predictive validity that test theorists always have wanted to convey, but does not depend on assumptions about correlated errors (Zimmerman, 1983, 1998).

## COMPONENTS OF SCORES IN CLASSICAL TEST THEORY

According to classical test theory, an observed score on a test is the sum of a true score and an error score, $X = T_X + E_X$. In the classical model, true and error scores are uncorrelated: the mean observed score equals the mean true score, the mean error score is zero, and the components satisfy many other algebraic identities. These familiar notions have been expressed in different ways by various authors (see, for example, Cronbach, Rajaratnam, & Gleser, 1963; Gulliksen, 1950; Guttman, 1945; Lord & Novick, 1968; Novick, 1966; Rozeboom, 1966; Steyer, 1988, 1989, 1990).

In formulations such as those of Guttman (1945), Lord and Novick (1968), and Novick (1966), a true score was defined as the expectation of an individual's observed scores over independent, repeated measurements or replications of a test. Lord and Novick introduced the expression "propensity distribution" and an accompanying notation to describe this hypothetical distribution of test scores, although in statistical theory such a distribution usually is implicit. Guttman (1945) had previously used a similar notation to make explicit the sources of variation in test scores and to make explicit probability distributions that mathematical statisticians usually regard as implicit.

From the definition of a true score as the expectation of an individual's observed scores, many familiar results can be derived without further assumptions. It is sometimes said that the classical model is obtained by construction, not by assumption (see, for example, Guttman, 1945; Novick, 1966). In this article, however, we discover that the simplicity of this model is somewhat deceptive. Our goal is to incorporate the theory in a more general mathematical framework, to eliminate remaining ambiguities, and to emphasize the relations of the basic concepts to other mathematical models. The point of view we are developing has been prominent in some areas of probability and statistics, but is not typically encountered in psychological measurement. We discover, however, that juxtaposition of the two perspectives yields insight into well-known results in test theory.

## PROBABILITY SPACES, TESTS, AND OBSERVED SCORES

We begin with a probability space $(\Omega,A,P)$, a random variable $X: \Omega \rightarrow R$ with finite variance, representing test scores, and a random point $f: \Omega \rightarrow \Phi$, where $\Phi$ is a set of individuals or experimental objects. The sample space $\Omega$ comprises all possible outcomes of a test procedure. The random variable X will be called the *observed score,* as is customary in test theory.

A random variable defined on a probability space is a familiar object in statistical theory. However, distinctive properties of test scores arise from the interrelation of the random variable X and the random point f. The function f can be regarded as an assignment of individuals or objects to sample points, which presupposes that one has selected a $\sigma$-algebra of subsets of $\Phi$ and that inverse images of sets in the collection belong to the $\sigma$-algebra A of subsets of $\Omega$. Usually, the set $\Phi$ is finite or countably infinite, representing a discrete population of individuals, so that the set of all subsets of $\Phi$ is a $\sigma$-algebra. If $\Phi$ is the set of real numbers, the Borel sets are taken as the σ-algebra. As in many probability models, a sample space often is implicit, and the primary objects of interest are random variables and probability distributions. These ideas are summarized by the following definition.

### Definition 1

A test is a 5-tuple (Ω,A,P,f,X), consisting of a set of outcomes Ω, a σ-algebra A of subsets of Ω, representing observable events, a set function P defined on A, and two point functions f and X defined on Ω, such that (Ω,A,P) is a probability space, f is a random point, and X is a random variable with finite variance.

Further semantic interpretation is as follows. An individual β is selected from a population Φ, and a measurement is taken, resulting in an observed score. The sample space Ω represents all conceivable outcomes of the entire procedure, which encompasses both individuals and measurements. The observed score random variable X is defined on Ω and represents possible scores of all individuals over the entire population. On the other hand, X restricted to a subset of Ω of the form $f^{-1}(\beta)$, that is, the conditional random variable X|f = β, represents the possible observed scores of individual β. The probability distribution of this conditional random variable (i.e., the propensity distribution) represents the inherent variability, or error of measurement, characterizing an individual's test score. In the case of an idealized perfectly reliable measurement, X|f = β is a constant function.

## THE CONCEPT OF TRUE SCORE

The concept of a true score is central in test theory and brings into prominence some probabilistic concepts not found in other statistical theories.

### Definition 2

The true score $T_X$ corresponding to the observed score X is a B-measurable random variable defined on the sample space Ω, such that

$$E[T_x|b] = E[X|b], \text{ for all } b \in B,$$

where E denotes expectation and $B \subseteq A$ is the σ-algebra of subsets of Ω induced by f.

According to this definition, the true score is a random variable defined on the same probability space as the observed score. The condition of B-measurability implies that $T_X$ is constant on subsets of Ω of the form $f^{-1}(\beta)$, the atoms of B, or, in other words, that the true score random variable assumes a constant value on subsets of Ω that correspond to particular individuals. In alternate language, definition 2 states that the true score is the conditional expectation of the random variable X given the σ-algebra induced by the random point representing individuals

(Zimmerman, 1975). Because the true score is itself a random variable defined on $\Omega$, it has a probability distribution, an expectation, variance, covariance with other random variables, and so on.

The expectation of the observed score equals the expectation of the true score, not only in the entire set $\Phi$, as in some versions of the classical theory, but also in any subpopulation of $\Phi$. Incorporating an entire collection of equalities into the definition, instead of deriving equalities one by one, has significant consequences that are also evident in the probabilistic theory of conditional expectation. It will become apparent in the pages to follow that definition 2 implies precisely those properties of true scores that are needed in test theory.

## Example 1

Table 1 describes a sample space $\Omega = \{\omega_1,\omega_2,\omega_3,\omega_4\}$, a discrete probability density p, a random point f, an observed score X, and a true score $T_X$. In this example, the $\sigma$-algebra induced by f is $B = \{\varnothing,\{\omega_1, \omega_2\},\{\omega_3, \omega_4\},\Omega\}$. Conditional expectations are $E[X \mid f = \beta_1] = .5$ and $E[X \mid f = \beta_2] = 1.5$. Therefore, $T_X(\omega_1) = T_X(\omega_2) = .5$, because $f(\omega_1) = f(\omega_2) = \beta_1$ and $T_X(\omega_3) = T_X(\omega_4) = 1.5$, because $f(\omega_3) = f(\omega_4) = \beta_2$. Note that $T_X$ is constant on inverse images of $\beta_1$ and $\beta_2$. The last column, representing an error score, will be explained later.

## Example 2

If a test is administered to a single individual, then $\Phi$ is a one-element set $\{\beta\}$. In this case, $B = \{\varnothing,\Omega\}$, and $T_X = EX$. None of the variation in the observed score X is accounted for by variation in values of $E[X \mid f = \beta]$.

## Example 3

Let f be a one-to-one function from $\Omega$ onto $\Phi$. Then, for all $\beta \in \Phi$, $f^{-1}(\beta)$ is a one-element subset of $\Omega$, so that B = A and $T_X = X$. This means that each individual's ob-

TABLE 1
Sample Space $\Omega = \{\omega_1,\omega_2,\omega_3,\omega_4\}$ With Observed Score Random Variable X, True Score $T_X$, and Error Score $E_X$

| | *p*( ) | *f*( ) | *X*( ) | $T_X$( ) | $E_X$( ) |
|---|---|---|---|---|---|
| $\omega_1$ | .25 | $\beta_1$ | 0 | 0.5 | –.5 |
| $\omega_2$ | .25 | $\beta_1$ | 1 | 0.5 | .5 |
| $\omega_3$ | .25 | $\beta_2$ | 1 | 1.5 | –.5 |
| $\omega_4$ | .25 | $\beta_2$ | 2 | 1.5 | .5 |

served score and true score are the same, or that all variation in the observed score X is accounted for by variation in the values of $E[\mathrm{X} \mid \mathrm{f} = \beta]$.

## THE TRUE SCORE AS A LINEAR OPERATOR ACTING ON RANDOM VARIABLES

At this point in the development of these notions, it is convenient to adopt a somewhat different perspective. We consider an operator, *T*, that assigns to each observed score random variable X its corresponding true score random variable $\mathrm{T_X}$, that is $T(\mathrm{X}) = \mathrm{T_X}$, or $\mathrm{X} \rightarrow \mathrm{T_X}$ under the operator *T*. The domain of the random variables X and $\mathrm{T_X}$ is the same sample space Ω, whereas the domain of the operator *T* is a collection of random variables, or functions. The codomain of X and $\mathrm{T_X}$ is R, the set of real numbers, whereas the codomain of the operator *T* is, again, a collection of random variables. The distinctive features of test theory as a mathematical model are closely related to the fact that *T* is a projection operator in Hilbert space. In the following pages, we represent random variables (or vectors) by ordinary Latin letters, possibly with subscripts, and operators by italic letters without subscripts.

Consider two observed score random variables X and Y defined on the same probability space. An important property of the operator *T* which accounts for its usefulness is that

$$T(\mathrm{c_1X} + \mathrm{c_2Y}) = \mathrm{c_1T_X} + \mathrm{c_2T_Y},$$

where $\mathrm{c_1}$ and $\mathrm{c_2}$ are arbitrary constants; that is, *T* is a linear operator. Given an observed score X, it is possible that another random variable Y is constant on each member of the partition of Ω induced by f, but is not the true score corresponding to X, because its constant value is not $E[X|\mathrm{B}]$. However, a random variable of this type possesses the convenient property that

$$\mathrm{T_{XY}} = \mathrm{T_XY}.$$

If Y is B-measurable, $\mathrm{T_Y} = \mathrm{Y}$. In particular, $T(\mathrm{T_X}) = \mathrm{T_X}$. That is, any random variable that is constant on sets of the form $\mathrm{b} \in \mathrm{B}$ is transformed into itself by the operator *T*, and a true score is transformed into the same true score. Finally, the familiar test–theory result $E[\mathrm{T_X}] = E[\mathrm{X}]$ follows, because $E[\mathrm{T_X}|\Omega] = E[\mathrm{X}|\Omega]$.

## TEST RELIABILITY

We now investigate covariances and correlations among observed score random variables and true score random variables. The equation

$$E[\mathrm{T_XY}] = E[\mathrm{XT_Y}] = E[\mathrm{T_XT_Y}],$$

which can be regarded either as a relation among operators or a relation among conditional expectations, provides a basis for numerous derivations relating true scores and error scores. For example, substituting X – *E*X for X and Y – *E*Y for Y produces the result

$$\mathrm{Cov}(T_X,Y) = \mathrm{Cov}(X,T_Y) = \mathrm{Cov}(T_X,T_Y).$$

In particular, $\mathrm{Cov}(X,T_X) = \mathrm{Cov}(T_X,T_Y) = \mathrm{Var}\ X$. From these identities it readily follows that the squared correlation between observed scores and true scores equals the ratio of true score variance and observed score variance, $\rho^2(X,T_X) = \mathrm{Var}\ T_X/\mathrm{Var}\ X$ , so that the ratio is a number between 0 and 1, and $\mathrm{Var}\ T_X \leq \mathrm{Var}\ X$. It is convenient to introduce the following definition.

### Definition 3

The reliability of a test $(\Omega,A,P,f,X)$, defined if X has nonzero variance, is the ratio $\mathrm{Var}\ T_X/\mathrm{Var}\ X$.

If X is normalized to have standard deviation 1, then reliability is $\mathrm{Var}\ T_X$. Reliability is 1 if and only if $\mathrm{Var}\ X = \mathrm{Var}\ T_X$ and is 0 if and only if $\mathrm{Var}\ T_X = 0$. In fact, a stronger statement is warranted: Reliability is 1 if and only if $X = T_X$.

Reliability has been defined in many different ways in test theory, and for most purposes it is immaterial which algebraic expression is taken as a definition and which expressions are regarded as theorems. In the discussion to follow we denote reliability by the commonly used symbol $\rho_{XX}$, which alludes to the fact that reliability equals a correlation between parallel tests. One slight advantage of taking the ratio of true score variance and observed score variance as the definition is that it encompasses all observed scores with nonzero variance, whereas the squared correlation $\rho^2(X,T_X)$ is not defined if true score variance is zero.

## THE ERROR SCORE AS A LINEAR OPERATOR

The concept of error score, like that of true score, will be introduced in a manner that first glance appears strange. Once again, however, we discover that this approach has decided advantages.

### Definition 4

The error score $E_X$ corresponding to the observed score X is a random variable defined on $\Omega$, such that

$$E[E_x|B] = 0, \text{ for all } b \in B,$$

where $B \subseteq A$ is the $\sigma$-algebra of subsets of $\Omega$ induced by f.

This definition should be compared to definition 2. It states that the mean error score is zero in the entire population of individuals, as well as any subpopulation of individuals. Setting B = Ω, we obtain $E[E_X] = 0$. From definition 4, along with definition 2, one derives the relation $X = T_X + E_X$.

Expressed in a more familiar way, the error score $E_X$, a random variable defined on the same probability space as X, is the pointwise difference of X and $T_X$. This concept is complementary to the conditional expectation, in the sense that the error score is the difference between an observed score and its conditional expectation given a σ-algebra. Because it is a random variable defined on Ω, the error score has a probability distribution, expectation, variance, and covariance with other random variables.

## Example 4

The last column in Table 1 shows an error score $E_X$ corresponding to the observed score X. Note that the sum of values of $E_X$ is 0, and the same is true of $E_X[f = \beta_1]$ and $E_X[f = \beta_2]$.

## Example 5

If f(Ω) is a one-element set {β}, so that $B = \{\varnothing, \Omega\}$ and $T_X = EX$, then $E_X = X - EX$, or the error score, is the same as the observed score centered at its expectation, or the deviation from the mean.

## Example 6

If f is one-to-one, so that B = A and $T_X = X$, then $X - T_X = 0$, or $E_X$ assumes the value 0 at each $\omega \in \Omega$.

It is convenient to introduce an operator *E*, analogous to the operator *T*, which transforms observed score random variables into error score random variables. This operator is defined by E = 1 – T, where 1 is the identity operator defined by $1(X) = X$, which transforms an observed score into itself. This notation is meaningful, because linear combinations of operators behave like linear combinations of functions. That is, $(1 - T)(X) = 1(X) - T(X) = X - T_X = E_X$. In operator notation, the familiar decomposition of observed scores into true and error components takes the form

$$1(X) = T(X) + E(X), \text{ for all X.}$$

## PROPERTIES OF ERROR SCORES

Many properties of *E* are similar to properties of *T*. First, $E(c_1X + c_2Y) = c_1E_X + c_2E_Y$, where $c_1$ and $c_2$ are constants. If Y is B-measurable, then, for any X, $E_{XY} = E_XY$, which corresponds to a similar result for $T_X$, and $E_Y = 0$, where 0 here denotes the function that takes the value 0 at each $\omega \in \Omega$. The operators *T* and *E* are related as follows: For all X, $T(T_X) = T_X$, $E(E_X) = E_X$, and $E(T_X) = T(E_X) = 0$. To jump ahead, the range of the operators *T* and *E*, denoted by *R*(T) and *R*(E), are as follows:

$$R(T) = \{X \mid T_x = X\} = \{X \mid E_x = 0\} \text{ and}$$
$$R(E) = \{X \mid T_x = 0\} = \{E_x = X\}.$$

We now investigate covariances and correlations among true scores and error scores and list a few formulas that are familiar in test theory. First,

$$E(E_X Y) = E(X E_Y) = E(E_X E_Y),$$

which is comparable to the previous result for *T*. Substituting X – *E*X for X and Y – *E*Y for Y, yields

$$\mathrm{Cov}(E_X,Y) = \mathrm{Cov}(X,E_Y) = \mathrm{Cov}(E_X,E_Y).$$

This result at first glance appears to be incongruous with classical test theory, which does not admit nonzero correlations of error scores and other variables. It is true that the initial definitions in the classical model, as well as the model in this article, imply that the covariance of any true score and any error score is zero. However, they do not imply that the covariance between two error scores is zero, and that result, if needed in a derivation, must be obtained by introducing an independent assumption.

The above equation relating covariances of error scores and observed scores makes it possible to derive additional equations containing correlations among error scores that are useful in some contexts. As one practical application of this more general point of view, some authors have pointed out that nonzero correlations among error components of scores on subtests of a composite test are likely to exist (Guttman, 1953; Rozeboom, 1966; Zimmerman & Williams, 1977, 1980).

Beginning with $E(XY) = E[(T_X + E_X)(T_Y + E_Y)]$, expanding, and using the fact that $E(T_X E_Y) = E(T_Y E_X) = 0$, yields $E(XY) = E(T_X T_Y) + E(E_X E_Y)$. Substituting X – *E*X for X and Y – *E*Y for Y, we obtain

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(T_X,T_Y) + \mathrm{Cov}(E_X,E_Y).$$

Setting X = Y immediately produces the result $\mathrm{Var}\ X = \mathrm{Var}\ T_X + \mathrm{Var}\ E_X$, and dividing by Var X the results $\rho_{XX'} = 1 - \mathrm{Var}\ E_X/\mathrm{Var}\ X$ and $\sigma_{E_X} = \sigma_X \sqrt{1-\rho_{XX'}}$. Although this derivation is straightforward and leads quickly to familiar results of the classical model, it makes no sense if one assumes at the outset that $\mathrm{Cov}(E_X,E_Y) = 0$ is an identity.

There are a large number of algebraic identities of this kind in the classical model, and we shall not dwell on these in this article. However, it is interesting to observe this fact: If *T* and *E* are interchanged and $\rho_{XX'}$ is replaced by $1 - \rho_{XX'}$, in any of these familiar identities in test theory, the result is another identity (Zimmerman, 1979). The significance of this duality in the model will become apparent in the following development.

## CONSTRUCTION OF PARALLEL MEASUREMENTS

Test theorists have proposed numerous definitions of parallel measurements, and under almost all definitions it turns out that the reliability coefficient equals the correlation between parallel measurements. It is generally agreed that the respective tests should possess the same true scores, the same observed score means, and the same observed score variances. It is also stipulated that error scores on parallel measurements are uncorrelated. Therefore, the two tests are interchangeable in the sense that they are equally effective as instruments despite the influence of error.

A strong definition, which reveals the relation of the concept to other statistical models, is that observed score random variables $X_1$ and $X_2$ are parallel, if, for each $\beta \in \Phi$, the restrictions of $X_1$ and $X_2$ to $f^{-1}(\beta)$ are independent, identically-distributed random variables. This point of view reveals that parallel measurements are constructed in somewhat the same sense that random samples are constructed in sampling theory. This construction implies that $X_1$ and $X_2$ are exchangeable, or symmetrically dependent. Otherwise stated, a sequence of n mutually parallel measurements can be regarded as a collection of random samples of size n parameterized by a set $\Phi$. All the usual test–theory results involving parallel tests can be derived from this definition.

A weaker definition, which can be conveniently expressed in the operator notation of this article and which still accomplishes what is intended in test theory, follows.

### Definition 5

Observed score random variables $X_1$ and $X_2$ are parallel if

$$T(X_1) = T(X_2),$$
$$T(X_1^2) = T(X_2^2),$$
$$\text{and } T(X_1X_2) = T(X_1)T(X_2).$$

From this definition it can be proved that parallel measurements have the same means and variance, as usually required, and that their restrictions to each individual are uncorrelated. Furthermore, if $X_1$ and $X_2$ are parallel, then, for all $\beta \in \Phi$, the conditional random variables $X_1 \mid f = \beta$ and $X_2 \mid f = \beta$ are parallel. In other words, as Lord and Novick (1968) emphasized, measurements that are parallel in a given population are also parallel in any subpopulation.

Under these definitions, it is easy to prove that the correlation between parallel measurements equals the ratio of the variance of true scores and the variance of observed scores. Therefore, it is immaterial whether one initially defines reliability as the correlation between parallel measurements or as a ratio of variances.

## TEST VALIDITY

We now investigate covariances and correlations among random variables that are not necessarily parallel according to the previous definitions. The concept of test validity to be introduced in this section diverges from conventional definitions and has some advantages that will soon appear (Zimmerman, 1983, 1998). Recall that the reliability of X is Var $T_X$/Var X and that the reliability of Y is Var $T_Y$/Var Y. The definition of test validity to be proposed possesses a certain symmetry when considered together with these formulas.

### Definition 6

Let X and Y be arbitrary observed score random variables with nonzero variance. The validity of X with respect to Y, denoted by $V_{XY}$, is the ratio $\text{Cov}(T_X,T_Y)/\sigma_X\sigma_Y$.

If X and Y are normalized variables, validity is just $\text{Cov}(T_X,T_Y)$. Obviously, the validity of X with respect to Y is the same as the validity of Y with respect to X. Traditionally, validity is the ordinary Pearson correlation between observed scores, $\rho(X,Y)$, or the absolute value of that correlation. From the relation $\text{Cov}(X,Y) = \text{Cov}(T_X,T_Y) + \text{Cov}(E_X,E_Y)$, together with definition 6, it follows that $V_{XY} = \rho(X,Y)$ if and only if $\rho(E_X,E_Y) = 0$. Under definition 6, however, it turns out that validity is a meaningful concept and possesses familiar properties even if errors are correlated.

As in the case of reliability, it is to some extent arbitrary which expression is taken as the definition of validity, and alternative derivations are possible. It should be emphasized, however, that the classical definition of validity as the correlation $\rho(X,Y)$ has been coupled with the assumption that $\rho(E_X,E_Y) = 0$, which introduces complications that did not arise when defining reliability.

### Example 7

Let the observed score random variables X and Y be defined as indicated in Table 2. The true and error scores corresponding to X and Y are shown. Calculation reveals that $\text{Cov}(X,Y) = .25$, Var X = Var Y = 1.25, $\text{Cov}(T_X,T_Y) = \text{Var } T_X = \text{Var } T_Y = .25$, and $\text{Cov}(E_X,E_Y) = 0$. The validity of X with respect to Y, according to definition 6, is $V_{XY} = \text{Cov}(T_X,T_Y)/\sigma_X\sigma_Y = .20$. Furthermore $\rho(X,Y) = \text{Cov}(X,Y)/\sigma_X\sigma_Y = .20 = V_{XY}$, so that definition 6 coincides with the classical definition in this case.

TABLE 2
Sample Space $\Omega = \{\omega_1,\omega_2,\omega_3,\omega_4\}$ With Observed Score Random Variables X and Y, True Scores $T_X$ and $T_Y$, and Uncorrelated Error Scores $E_X$ and $E_Y$

| | $p(\;)$ | $f(\;)$ | $X(\;)$ | $Y(\;)$ | $T_X(\;)$ | $T_Y(\;)$ | $E_X(\;)$ | $E_Y(\;)$ |
|---|---|---|---|---|---|---|---|---|
| $\omega_1$ | .25 | $\beta_1$ | 2 | 4 | 1 | 3 | 1 | 1 |
| $\omega_2$ | .25 | $\beta_1$ | 0 | 2 | 1 | 3 | –1 | –1 |
| $\omega_3$ | .25 | $\beta_2$ | 1 | 5 | 2 | 4 | –1 | 1 |
| $\omega_4$ | .25 | $\beta_2$ | 3 | 3 | 2 | 4 | 1 | –1 |

## Example 8

Let the observed score random variables X and Y be defined as indicated in Table 3. We find that Cov(X,Y) = –.75, Var X = Var Y = 1.25, Cov($T_X$,$T_Y$) = Var $T_X$ = Var $T_Y$ = .25, and Cov($E_X$,$E_Y$) = -1.00. The validity of X with respect to Y by definition 6 is, $V_{XY}$ = .20, but ρ(X,Y) = –.60.

The reliability of a test, we have seen, can be regarded as a correlation between parallel measurements. Under definition 6, a similar property characterizes validity (Zimmerman, 1983, 1998). Beginning with random variables X and Y, it is possible to construct random variables X* and Y*, which satisfy the following equalities:

$$\begin{gathered} T(\mathrm{X}^*) = T(\mathrm{X}) \\ T(\mathrm{X}^{*2}) = T(\mathrm{X}^2) \\ T(\mathrm{Y}^*) = T(\mathrm{Y}) \\ T(\mathrm{Y}^{*2}) = T(\mathrm{Y}^2) \\ \text{and } T(\mathrm{X}^*\mathrm{Y}^*) = T(\mathrm{XY}) = T(\mathrm{X}^*)T(\mathrm{Y}^*). \end{gathered}$$

These random variables X* and Y* are dependent, but their restrictions to $f^{-1}(\beta)$, for each $\beta \in \Phi$, are uncorrelated, so this concept is a generalization of parallel measurements. If X* and Y* possess these properties in a population, then they possess the same properties in any subpopulation. It is possible to prove that

$$V_{XY} = \rho(X^*,Y^*),$$

whatever the correlation ρ($E_X$,$E_Y$) may be (Zimmerman, 1983).

In probability theory, if one is given a random variable X with an arbitrary probability distribution, it is always possible to construct two random variables $X_1$ and $X_2$ which are independent and identically distributed. These random variables are defined on the Cartesian product of two copies of the original sample space, and probabilities are assigned in the product space according to a product rule. This construction sometimes is called an independent subexperiments model.

TABLE 3
Sample Space $\Omega = \{\omega_1,\omega_2,\omega_3,\omega_4\}$ With Observed Score Random Variables X and Y, True Scores $T_X$ and $T_Y$, and Correlated Error Scores $E_X$ and $E_Y$

| ω | *p*( ) | *f*( ) | *X*( ) | *Y*( ) | $T_X$( ) | $T_Y$( ) | $E_X$( ) | $E_Y$( ) |
|---|---|---|---|---|---|---|---|---|
| $\omega_1$ | .25 | $\beta_1$ | 2 | 4 | 1 | 3 | 1 | –1 |
| $\omega_2$ | .25 | $\beta_1$ | 0 | 2 | 1 | 3 | –1 | 1 |
| $\omega_3$ | .25 | $\beta_2$ | 1 | 5 | 2 | 4 | 1 | –1 |
| $\omega_4$ | .25 | $\beta_2$ | 3 | 3 | 2 | 4 | –1 | 1 |

Similarly, given an observed score X, it is always possible by means of a similar construction to obtain random variables $X_1$ and $X_{2,}$ which are parallel according to definition 5. Finally, given arbitrary random variables X and Y, it is possible to construct random variables X* and Y*, which satisfy the above generalization of parallel measurements. Roughly, the correlation between X* and Y* is the correlation between X and Y, which would be obtained if statistical dependence among errors were eliminated.

An advantage of this approach is that some familiar equations, including the Spearman attenuation formula, can be derived in their usual form despite the existence of correlated errors. That is,

$$\rho\left(T_X, T_Y\right) = \frac{V_{XY}}{\sqrt{\rho_{XX'}\rho_{YY'}}},$$

whatever the correlation between error scores $E_X$ and $E_Y$ may be. However, ρ(X,Y) cannot be substituted for $V_{XY}$ in this formula unless $\rho(E_X,E_Y) = 0$. This is apparent from the relation

$$\rho(X,Y) = V_{XY} + \rho\left(E_X, E_Y\right)\sqrt{\left(1-\rho_{XX'}\right)\left(1-\rho_{YY'}\right)}.$$

It is also true that $V_{XY} \leq \sqrt{\rho_{XX'}\rho_{YY'}}$. Therefore, the index of reliability, $\sqrt{\rho_{XX'}}$, is always an upper bound on validity, as defined in this article, but not on ρ(X,Y) unless $\rho(E_X,E_Y) = 0$.

The extent to which $V_{XY}$ and ρ(X,Y) differ depends on the reliability coefficients of the respective tests. The discrepancy can be a serious problem when employing the Spearman attenuation formula if these reliability coefficients are small. The discrepancy is

$$\frac{\rho\left(E_X, E_Y\right)\sqrt{\left(1-\rho_{XX'}\right)\left(1-\rho_{YY'}\right)}}{\sqrt{\rho_{XX'}\rho_{YY'}}},$$

which increases without bound as $\rho_{xx'}$ and $\rho_{yy'}$ become small.

Correlated error scores can be considered from another point of view. Let $\Psi_X$ be a function that assigns to each individual β the variance of X | f = β, that is, the variance of the individual's scores over independent, repeated measurements, and let $\Psi_Y$ be a similar function. Guttman(1945) defined test reliability as $1 - E\Psi_X/\mathrm{Var}\ X$, where the expectation is taken over individuals. Furthermore, let $\Psi_{XY}$ denote a function that assigns to each individual β the covariance between X | f = β and Y | f = β. Then the correlation between error scores is given by the following equation:

$$\rho\left(E_X E_Y\right) = \frac{E\psi_{XY}}{\sqrt{E\psi_X E\psi_Y}}.$$

Guttman (1953) and Rozeboom (1966) expressed dissatisfaction with assumptions about correlated errors in test theory. Interestingly, the aforementioned equation gives explicitly the magnitude of something that was generally assumed not to exist at the time these authors expressed their concerns. See also Zimmerman and Williams (1977, 1980).

## FUNCTION SPACES, METRIC SPACES, AND HILBERT SPACES

Psychologists are familiar with various statistical models in which random variables are regarded as vectors, and matrix representations are useful in interpreting properties of variance, standard deviation, covariance, and correlation in the models. Because random variables are real-valued functions, they may be identified with points or vectors in a vector space of functions. The values assumed by a random variable at each point ω of a sample space Ω are the coordinates of a vector relative to a basis. This approach is fundamental in modern multivariate statistics, although apart from the matrix symbolism, the abstract formalism of vector spaces often remains in the background.

In test theory, the collection of all observed score random variables defined on a probability space associated with a test procedure (see definition 1) is one example of a vector space of random variables. Because observed scores have finite variance, an inner product can be introduced into this space in such a way that it becomes a complete inner product space or Hilbert space (see, for example, Burrill, 1972; Edwards, 1965; Halmos, 1972; Loève, 1963; Rényi, 1970). Metric concepts such as length, distance, angle, and orthogonality represent properties of observed scores that are familiar in test theory.

Furthermore, the collection of all true score random variables, or B-measurable random variables, defined on the same probability space, is a Hilbert subspace of the space of observed score random variables. The operators *T* and *E* that have been prominent in this article are orthogonal projections onto that subspace and its orthogonal complement, respectively. Once these identifications have been made, it is possible to interpret many familiar results in test theory in geometric terms. In fact, proofs of test theory results can be obtained directly from metric properties of the Hilbert space. Moreover, this approach yields insight into the structure of the theory as a whole and shows that several definitions in this article that at first glance appear anomalous are in reality quite natural.

## THE HILBERT SPACE $L_2(\Omega,A,P)$

The collection of all random variables defined on a given probability space $(\Omega,A,P)$ is a vector space over the field of real numbers, typical of function spaces investigated in functional analysis. The values assumed by the random variable, or the real numbers that are images of sample points, are the coordinates of the vector relative to a basis. Collections of random variables defined on finite sample spaces are finite-dimensional vector spaces. Collections defined on countably infinite or uncountable sample spaces, which are of most interest for our present purposes, are infinite-dimensional vector spaces.

To interpret concepts in probability, statistics, and test theory, it is not necessary to consider the collection of all random variables defined on a probability space. It is sufficient to restrict attention to the collection of all random variables having finite variance, or, as sometimes called, square-integrable random variables. Because random variables with finite variances also possess finite covariances and expectations, this collection is sufficiently large to provide for an interpretation of test theory.

It is possible to introduce an inner product into this vector space of random variables and, in turn, to derive a norm and a metric from the inner product. It urns out that the space is complete with respect to the metric derived from the inner product, so this particular vector space is a Hilbert space. All the metric concepts familiar in the Euclidean spaces $E^2$ and $E^3$, including length, distance, angle, and orthogonality, are meaningful.

The inner product which imparts the structure of a Hilbert space to this collection of random variables is $E(XY)$, the expectation of the product of the random variables. The norm determined by this inner product is $\sqrt{EX^2}$, and the metric is $\sqrt{E(X-Y)^2}$. If attention is restricted to random variables centered at expectations—that is, to deviations from the mean—then the norm, or length, of a random variable is the standard deviation, the cosine of the angle between two random variables is the correlation coefficient, and two random variables are orthogonal if and only if they are uncorrelated. The Hilbert space of random variables with finite variance defined on a probability space $(\Omega,A,P)$, having this structure, is customarily denoted by $L_2(\Omega,A,P)$, or sometimes simply by $L_2(A)$, or by $L_2(P)$.

## THE HILBERT SUBSPACE $L_2(\Omega,B,P)$

Consider now the collection of all true score random variables corresponding to observed score random variables in $L_2(\Omega,A,P)$. This collection of B-measurable random variables will be denoted by $L_2(\Omega,B,P)$, or simply by $L_2(B)$. A member of this collection is transformed into itself by the operator *T*. That is,

$$L_2(\Omega,B,P) = L_2(B) = \{X \in L_2(A) \mid T(X) = X\}.$$

This collection of true score random variables is a closed vector subspace, or Hilbert subspace of $L_2(a)$, which possesses the same metric structure, including lengths, angles, and orthogonality.

The orthogonal complement of $L_2(B)$, which will be denoted by $L_2(B)^\perp$, is the collection of all random variables in $L_2(A)$ having the property of being orthogonal to every random variable in $L_2(B)$. The test–theory model is based on the following two identifications: The true score $T_X$ corresponding to an observed score $X \in L_2(A)$ is the orthogonal projection of X onto $L_2(B)$, and the error score $E_X$ corresponding to X is the orthogonal projection of X onto $L_2(B)^\perp$. That is, the linear operators *T* and *E* are orthogonal projections.

Accordingly, the collection of all error score random variables can be identified with the orthogonal complement of the collection of all true score random variables, and every observed score $X \in L_2(A)$ is a unique sum of a true score and an error score, $X = T_X + E_X$, where $T_X \in L_2(B)$ and $E_X \in L_2(B)^\perp$. Therefore, true scores and error scores are orthogonal, or uncorrelated. Any random variable in the subspace of all true scores is orthogonal to any random variable in the subspace of all error scores. However, two arbitrary random variables $E_X$ and $E_Y$ in the subspace of error scores are not necessarily uncorrelated, just as arbitrary $T_X$ and $T_Y$ are not necessarily uncorrelated.

## METRIC INTERPRETATION OF RELIABILITY AND VALIDITY

Now that statistical concepts of variance, covariance, and correlation have been related to metric concepts in $L_2(A)$ and true scores and error scores have been identified with orthogonal projections onto the subspaces $L_2(B)$ and $L_2(B)^\perp$, it is possible to provide geometric interpretations of reliability, validity, and other test–theory concepts. The algebraic equations relating variances, covariances, and correlations that are familiar in the classical model emerge from the formalism in a natural way. Furthermore, the Hilbert space formalism brings to light some relations that are not obvious in the classical model and yields further insight into the structure of the theory.

Orthogonal projections are idempotent and self-adjoint, which in this context means that $T(T_X) = T_X$ and that $E[T(X)Y] = E[XT(Y)]= E(T_XY) = E(XT_Y)$, for all

$X,Y \in L_2(A)$. Considering $X - EX$ and $Y - EY$, the latter equality becomes $\mathrm{Cov}(T_XY) = \mathrm{Cov}(X,T_Y)$. In particular, $\mathrm{Cov}(T_XX) = \mathrm{Cov}(T_X,T_X) = \mathrm{Var}\ T_X$. The operator $E$ also is idempotent and self-adjoint, so various properties of true scores have analogues for error scores which are not evident from the classical definition of error score. Equalities corresponding to the ones previously stated for true scores are $E(E_X) = E_X$, $\mathrm{Cov}(E_XY) = \mathrm{Cov}(X,E_Y)$ and $\mathrm{Cov}(E_X,X) = \mathrm{Var}\ E_X$.

The square-root of the reliability coefficient, $\sigma(T_X)/\sigma_X$, is the ratio of the length of the projection of $X - EX$ onto the subspace of true scores to the length of $X - EX$. This ratio is a number between 0 and 1, which means that the norm of an orthogonal projection operator in the Hilbert space of all operators on a given Hilbert space is 1. The reliability coefficient is also the squared correlation between observed scores and true scores, which is the same as the squared cosine of the angle between $X - EX$ and its orthogonal projection.

Expressed in another way, a reliable test score is one which is “close” to the subspace of true scores, so that the length of its projection is almost the same as its own length. Such ideas are familiar in least-squares regression. If the length of the projection is decidedly less than that of the original vector, the two are “almost” perpendicular, so that reliability is close to zero. Along the same lines, the reliability of a test can be regarded as the “Rayleigh quotient” of an observed score centered at its expectation with respect to the true score operator (see, for example, Kreyszig, 1978, p. 469).

Many familiar algebraic formulas relating true scores, error scores, and observed scores that have been prominent in classical test theory are essentially trigonometric identities relating lengths to sines, cosines, and tangents of angles (Jackson, 1924) . For example, the fundamental equation $\mathrm{Var}\ X = \mathrm{Var}\ T_X + \mathrm{Var}\ E_X$ is the Pythagorean relation. The formula for the standard error of measurement, $\sigma_{E_X} = \sigma_X \sqrt{(1 - \rho_{XX'})}$ states that the length of $E_X$ is the product of the length of $X - EX$ and the cosine of the angle between $X - EX$ and $E_X$. These generalized lengths of random variables and angles between random variables possess some, but not all, of the properties of lines and angles in three-dimensional Euclidean space. This geometric interpretation of test–theory concepts encompasses parallel tests, composite tests, validity, and many other areas, but we shall not pursue these matters in this article. The geometric interpretation also provides insight into why difference scores and gain scores have led to complications in test theory (Zimmerman, 1997).

## THE SIGNIFICANCE OF ALTERNATIVE FORMULATIONS OF TEST THEORY

In the theory of stochastic processes, second order properties of random variables are those properties that depend only on the existence of variances and covariances.

Investigations sometimes establish relations among variances and covariances that do not involve details of the probability distributions or joint probability distributions of random variables (Loève, 1963). For example, random variables that are uncorrelated, but not necessarily independent, are prominent in these investigations.

The appropriate formalism underlying this approach is the Hilbert space $L_2(A)$, where metric relations derived from an inner product correspond to variances and covariances, as discussed previously, and where orthogonality assumes the role of independence. From this perspective, traditional models in test theory describe second order properties of random variables representing test scores. The familiar algebraic formulas in the classical theory, we have seen, can be interpreted as geometric relations in the space $L_2(A)$, and some of these relations are just trigonometric identities relating generalized lengths and angles. Of course, this fact does not prevent one from making a more detailed study of the probability distributions and joint probability distributions of test scores when desired, and stronger models frequently are informative.

The distinguishing feature of test theory, which sets it apart from other models in multivariate statistics, is the identification of true scores with elements of the subspace $L_2(B)$ and the orthogonal projection operators *T* and *E*, which have been discussed in this article. That is, the concepts of true score and error score bestow on test theory attributes not shared by other statistical models which also have as their appropriate formalism the space $L_2(A)$. Briefly expressed, classical test theory investigates second order properties of random variables that can be decomposed into true and error components.

Classical theory as ordinarily formulated also is restricted because it investigates only random variables with uncorrelated errors—that is, vectors in $L_2(A)$ having orthogonal projections onto $L_2(B)^{\perp}$ that are themselves orthogonal. This restriction is independent of the orthogonality of true scores and error scores, which is a property of projections onto complementary subspaces. Vectors within each of these subspaces are not necessarily orthogonal to each other.

It is possible to dispense with the restriction to random variables with uncorrelated errors and to derive results which encompass the entire space $L_2(A)$. This approach was taken in this article; that is, no assumptions were made about correlations among errors in the definitions of true scores, error scores, reliability, validity, and so on. Having derived a general formula that allows correlated errors, it is always possible to recover more familiar results by setting $\mathrm{Cov}(E_X,E_Y) = 0$. This is unnecessary in the context of reliability, when a single test score X is considered, but in the context of validity, when scores X and Y are investigated, or in the case of scores on composite tests, which are a sum of subtest scores, covariances between error scores sometimes exist.

The main results in test theory can be derived in several different ways, just as theorems in many branches of mathematics can be derived from alternate sets of axioms. In a mathematical system, it is often largely a matter of taste which statements are taken to be “axioms” and which are taken to be “theorems.” The same flexibility also characterizes “definitions.” A simple example is the several equivalent definitions of test reliability mentioned in this article. If reliability is first defined as the correlation between parallel measurements, it can be proved that it equals the ratio Var $T_X$/Var X. Alternatively, if reliability is initially defined as this ratio, it can be proved to be the correlation between parallel measurements, or the squared correlation between observed scores and true scores, and so on.

Furthermore, the basic concepts of test theory can be discussed on several different levels, and coordinating definitions which link the formalism to empirical terms can be introduced in alternate ways. An approach that differs from the one in this article begins with the theory of conditional expectation. A true score is defined as the conditional expectation of a random variable representing an observed score given the $\sigma$-algebra induced by a random point representing individuals (see, for example, Steyer, 1988, 1989; Zimmerman, 1975, for further discussion of this point of view). The results of this approach are equivalent to those in this article.

It is also possible to identify test scores with Hilbert space vectors initially, to provide metric interpretations of variance, covariance, and correlation, to identify true scores and error scores with orthogonal projections, and finally to derive the principal test–theory results in very few steps. In other words, it is logically acceptable to reverse the progression in this article and to obtain test theory directly from Hilbert space concepts at an abstract level. Even at this level, there are different ways to proceed. The language of linear operators was adopted in this article, but it is also possible to restrict attention to finite-dimensional vector spaces and to employ the language of matrices as commonly done in multivariate statistics. As noted earlier, this state of affairs is similar to that which prevails in quantum mechanics, where matrix mechanics and wave mechanics are formally equivalent, and in other mathematical models.

In these alternative formulations of test theory, it is necessary at some stage to introduce coordinating definitions which identify test scores with random variables, or vectors, and true scores with conditional expectations, or orthogonal projections. These definitions are simple and few in number, so that the theory in its entirety can be incorporated into a more general mathematical context all at once, and it is not necessary to make new assumptions continually as derivations proceed. In these respects, the formalism of test theory is quite similar to that of other scientific theories that are supported by a highly developed mathematical structure.

## ACKNOWLEDGMENT

This article was already targeted for this journal when Bruno D. Zumbo accepted the editorship of the *International Journal of Testing.* Therefore, as standard practice, the reviews were considered, and an independent party made the editorial decision.

## REFERENCES

Burrill, C. W. (1972). *Measure, integration, and probability.* New York: McGraw-Hill.

Cohen, D. W. (1989). *An introduction to Hilbert space and quantum logic*. New York: Springer-Verlag.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137–163.

Doob, J. L. (1953). *Stochastic processes.* New York: Wiley.

Edwards, R. E. (1965). *Functional analysis.* New York: Holt, Rinehart & Winston.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

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

Guttman, L. (1953). Reliability formulas that do not assume experimental independence. *Psychometrika, 18,* 123–130.

Halmos, P. (1972). *Introduction to Hilbert space.* New York: Chelsea.

Jackson, D. (1924). The trigonometry of correlation. *American Mathematical Monthly, 31,* 275–280.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung* [Foundations of Probability]. Berlin, Germany: Springer-Verlag.

Kreyszig, E. (1978). *Introductory functional analysis with applications.* New York: Wiley.

Loève, M. (1963). *Probability theory.* New York: Van Nostrand.

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

Messiah, A. (1959). *Quantum mechanics.* New York: Wiley.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1–18.

Rényi, A. (1970). *Foundations of probability.* San Francisco: Holden-Day.

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

Steyer, R. (1988). Conditional expectations: An introduction to the concept and its applications in empirical sciences. *Methodika, 2,* 53–78.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika, 3,* 25–60.

Steyer, R., & Schmitt, M. (1990). The effects of aggregation across and within occasions on consistency, specificity, and reliability. *Methodika, 4,* 58–94.

Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika, 40,* 395–412.

Zimmerman, D. W. (1979). A simple duality principle in test theory. *Journal of Mathematical Psychology, 20,* 256–262.

Zimmerman, D. W. (1983). The mathematical definition of test validity. *Educational and Psychological Measurement, 43,* 791–796.

Zimmerman, D. W. (1997). A geometric interpretation of the validity and reliability of difference scores. *British Journal of Mathematical and Statistical Psychology, 50,* 73–80.

Zimmerman, D. W. (1998). How should classical test theory have defined validity? Bruno D. Zumbo (Ed.), *Social indicators research, 45,* 233–251.

Zimmerman, D. W., & Williams, R. H. (1977). The theory of test validity and correlated errors of measurement. *Journal of Mathematical Psychology, 16,* 135–152.

Zimmerman, D. W., & Williams, R. H. (1980). Is classical test theory “robust” under violation of the assumption of uncorrelated errors? *Canadian Journal of Psychology, 34,* 227–237.