

- Flash T, Hogan N 1985 The coordination of arm movements: An experimentally confirmed mathematical model. *Journal of Neuroscience* 5: 1688–703
- Goldstein H 1980 *Classical Mechanics*. Addison–Wesley, Reading, MA
- Hollerbach J M, Flash T 1982 Dynamic interactions between limb segments during planar arm movement. *Biological Cybernetics* 44: 67–77
- Kawato M, Wolpert D 1998 Internal models for motor control. *Novartis Foundation Symposium* 218: 291–307
- Marr D 1982 *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W H Freeman and Co, San Francisco, CA
- Morasso P 1981 Spatial control of arm movements. *Experimental Brain Research* 42: 223–7
- Sabes P N, Jordan M I, Wolpert D M 1998 The role of inertial sensitivity in motor planning. *Journal of Neuroscience* 18: 5948–57
- Schaal S 1999 Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences* 3: 233–42
- Shadmehr R, Mussa-Ivaldi F A 1994 Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience* 14: 3208–24

S. Mussa-Ivaldi

## Classical (Psychometric) Test Theory

### 1. Introduction

One of the most striking and challenging phenomena in the Social Sciences is the unreliability of its measurements: Measuring the same attribute twice often yields two different results. If the same measurement instrument is applied twice, such a difference may sometimes be due to a change in the measured attribute itself. Sometimes these changes in the measured attribute are due to the mere fact of measuring. For example, people learn when solving tasks and they change their attitude when they reflect on statements in an attitude questionnaire. In other cases the change of the measured attribute is due to developmental phenomena, or it might be due to learning between occasions of measurement. However, if change of the attribute can be excluded two different results in measuring the same attribute can be explained only by 'measurement error.'

Classical (Psychometric) Test Theory (CTT) aims at studying the reliability of a (real-valued) test score variable (measurement, test) that maps a crucial aspect of qualitative or quantitative observations into the set of real numbers. Aside from determining the reliability of a test score variable itself, CTT allows answering questions such as:

(a) How do two random variables correlate once the measurement error is filtered out (correction for attenuation)?

(b) How dependable is a measurement in characterizing an attribute of an individual unit, i.e., which is the confidence interval for the true score of that individual with respect to the measurement considered?

(c) How reliable is an aggregated measurement consisting of the average (or sum) of several measurements of the same unit or object (Spearman–Brown formula for test length)?

(d) How reliable is a difference, e.g., between a pre-test and post-test?

## 2. Basic Concepts of Classical Test Theory

### 2.1 Primitives

In the framework of CTT, each measurement (test score) is considered being a value of a random variable  $Y$  consisting of two components: a 'true score' and an 'error score.' Two levels, or more precisely, two random experiments may be distinguished: (a) sampling an observational unit (e.g., a person) and (b) sampling a score within a given unit. Within a given unit, the true score is a parameter, i.e., a given but unknown number characterizing the attribute of the unit, whereas the error is a random variable with an unknown distribution. The true score of the unit is defined to be the expectation of this intraindividual distribution.

Taking the across units perspective, i.e., joining the two random experiments, the true score is itself considered to be a value of a random variable (the 'true score variable'). The 'error variable' is again a random variable, the distribution of which is a mixture of the individual units' error distributions. Most theorems of CTT (e.g., Lord and Novick, 1968) are formulated from this across units' perspective allowing talking about the correlation of true scores with other variables, for instance.

More formally, CTT refers to a (joint) random experiment of (a) sampling an observational unit  $u$  (such as a person) from a set  $\Omega_v$  of units (called the population), and (b) registering one or more observations out of a set  $\Omega_o$  of possible observations. The set of possible outcomes of the random experiment is the set product:  $\Omega = \Omega_v \times \Omega_o$ . The elements of  $\Omega_o$ , the observations, might be qualitative (such as 'answering in category  $a$  of item 1 and in category  $b$  of item 2'), quantitative (such as reaction time and alcohol concentration in the blood), or consisting of both qualitative and quantitative components. In Psychology, the measurements are often defined by test scoring rules prescribing how the observations are transformed into test scores. (Hence, these measurement are also often called 'tests' or 'test score variables.') These scoring rules may just consist of summing initial scores of items (defining a psychological scale) or

1955

**Table 1**  
Basic Concepts of Classical Test Theory

Primitives	
The set of possible events of the random experiment	$\Omega = \Omega_U \times \Omega_o$
Test Score Variables	$Y_i: \Omega \rightarrow \mathbb{R}$
Projection	$U: \Omega \rightarrow \Omega_U$
Definition of the Theoretical Variables	
True Score Variable	$\tau_i := E(Y_i U)$
Measurement Error Variable	$\varepsilon_i := Y_i - \tau_i$

might be more sophisticated representations of observable attributes of the units. CTT does not prescribe the definition of the test score variables. It just additively decomposes them into true score variables and error variables. Substantive theory and empirical validation studies are necessary in order to decide whether or not a given test score variable is meaningful. CTT only helps disentangling the variances of its true score and error components.

Referring to the joint random experiment described above the mapping  $U: \Omega \rightarrow \Omega_U, U(\omega) = u$ , (the unit or person projection) may be considered a qualitative random variable having a joint distribution with the test scores variables  $Y_i$ . Most theorems of CTT deal with two or more test score variables (tests)  $Y_i$  and the relationship between their true score and error components. (The index  $i$  refer to one of several tests considered.)

2.2 The core concepts: True score and error variables

Using the primitives introduced above, the true score variable  $\tau_i := E(Y_i|U)$  is defined by the conditional expectation of the test  $Y_i$  given the variable  $U$ . The values of the 'true score variable'  $\tau_i$  are the conditional expected values  $E(Y_i|U = u)$  of  $Y_i$  given the unit  $u$ . They are also called the 'true scores' of the unit  $u$  with respect to  $Y_i$ . Hence, these true scores are the expected values of the intraindividual distributions of the  $Y_i$ . The 'measurement error variables'  $\varepsilon_i$  are simply defined by the difference  $\varepsilon_i := Y_i - \tau_i$ . Table 1 summarizes the primitives and definitions of the basic concepts of CTT.

3. Properties of True Score and Error Variables

Once the true score variables and error variables are defined a number of properties (see Table 2) can be derived, some of which are known as the 'axioms of CTT.' However, since the work done by Novick (1966) and Zimmerman (1975, 1976) it is well known that all these properties already follow from the definition of true score and error variables. They are not new and independent assumptions as has been originally proposed (e.g., Gulliksen 1950). All equations in Table 2 are no assumptions. They are inherent properties of

true scores and errors. Hence, trying to test or falsify these properties empirically would be meaningless in just the same way, as it is meaningless to test whether or not a bachelor is really unmarried. The property of being unmarried is an inherent part or logical consequence of the concept of a bachelor.

Only one of the 'axioms of CTT' does not follow from the definition of true score and error variables: 'uncorrelatedness of errors variables' among each other. Hence, uncorrelatedness of errors has another epistemological status as the properties displayed in Table 2 (the other 'axioms'). Uncorrelatedness of errors is certainly a desirable and useful property; but it might be wrong in specific empirical applications (e.g., Zimmerman and Williams 1977). In fact it is an assumption and it plays a crucial rule in defining models of CTT.

Equation 1 of Table 2 is a simple rearrangement of the definition of the error variable. Equation. 2 shows that the variance of a test score variable, too, has two additive components: the 'variance of the true score variable' and the 'variance of the error variable.' This second property follows from Eqn. 3 according to which a true score variable is uncorrelated with a measurement error variable, even if they pertain to different test score variables  $Y_i$  and  $Y_j$ . Equation 4 states that the expected value of an error variable is zero, whereas Eqn. 5 implies that the expected value of an error variable is zero within each individual observational unit  $u$ . Finally, according to Eqn. 6 the conditional expectation of an error variable is also zero for each mapping of  $U$ . This basically means that the expected value of an error variable is zero in each subpopulation of observational units.

3.1 Additional Concepts: Reliability, Unconditional and Conditional Error Variances

Although the true score and error variables defined above are the core concepts of CTT, in empirical applications, the true scores can only be estimated. What is also possible, is to estimate the 'variances' of the true score and error variables in a random sample (consisting of repeating many times the random experiment described earlier). The variance  $Var(\varepsilon_i)$  of the measurement error may be considered a gross parameter representing the degree of unreliability. A normed parameter of unreliability is  $Var(\varepsilon_i)/Var(Y_i)$ , the proportion of the variance of  $Y_i$  due to measurement error. Its counterpart is  $1 - Var(\varepsilon_i)/Var(Y_i)$ , i.e.,

$$Rel(Y_i) := Var(\tau_i)/Var(Y_i) \tag{1}$$

the 'reliability' of  $Y_i$ . This coefficient varies between zero and one. In fact, most theorems and most empirical research deal with this 'coefficient of reliability.' The reliability coefficient is a convenient information about the dependability of the measurement 'in one single number.'

**Table 2**  
Properties of True Score and Error Variables Implied by Their Definition

Decomposition of the Variables	$Y_i = \tau_i + e_i$	(1)
Decomposition of the Variances	$Var(Y_i) = Var(\tau_i) + Var(e_i)$	(2)
Other Properties of True Score and Error Variables implied by their definition	$Cov(\tau_i, e_j) = 0$	(3)
	$E(e_i) = 0$	(4)
	$E(e_i U) = 0$	(5)
	for each (measurable) mapping of $U$ : $E[e_i f(U)] = 0$	(6)

In early papers on CTT, reliability of a test has been defined by its correlation with itself (e.g., Thurstone 1931, p. 3). However, this definition is only metaphoric, because a variable always correlates perfectly with itself. What is meant is to define reliability by the correlation of 'parallel tests' (see below). The assumptions defining parallel tests in fact imply that the correlation between two test score variables is the reliability. Note that the definition of 'reliability' via Eqn. (1) does not rest on any assumption other than  $0 < Var(Y_i) < \infty$ .

'Reliability' is useful to compare different instruments to each other if they are applied in the same population. Used in this way, reliability in fact helps evaluating the quality of measurement instruments. However, it may not be useful under all circumstances to infer the dependability of measures of an individual unit. For the latter purpose one might rather look at the 'conditional error variance'  $Var(e_i|U = u)$  given a specific observational unit  $u$  or at the 'conditional error variances'  $Var(e_i|\tau_i = t)$  given the sub-population with true score  $\tau_i = t$ .

4. Models of Classical Test Theory

The definitions of true score and error variables have to be supplemented by assumptions defining a model if the theoretical parameters such as the reliability are to be computed by empirically estimable parameters such as the means, variances, covariance, or correlation of the test score variables. Table 3 displays the most important of these assumptions and the most important models defined by combining some of these assumptions.

**Table 3**  
Assumptions and Some Models of CTT

Assumption used to define some models of CTT	
(a <sub>1</sub> ) $\tau$ -equivalence	$\tau_i = \tau_j$
(a <sub>2</sub> ) essential $\tau$ -equivalence	$\tau_i = \tau_j + \lambda_{ij}, \lambda_{ij} \in \mathbb{R}$
(a <sub>3</sub> ) $\tau$ -congenerity	$\tau_i = \lambda_{ij0} + \lambda_{ij1}\tau_j, \lambda_{ij0}, \lambda_{ij1} \in \mathbb{R}, \lambda_{ij1} > 0$
(b) uncorrelated errors	$Cov(e_i, e_j) = 0, i \neq j$
(c) equal error variances	$Var(e_i) = Var(e_j)$
Models defined by combining these assumptions	
<i>Parallel tests</i> are defined by Assumptions (a <sub>1</sub> ), (b) and (c).	
<i>Essentially <math>\tau</math>-equivalent tests</i> are defined by Assumptions (a <sub>2</sub> ) and (b).	
<i>Congeneric tests</i> are defined by Assumptions (a <sub>3</sub> ) and (b).	

Note: The equations refer to each pair of tests  $Y_i$  and  $Y_j$  of a set of tests  $Y_1, \dots, Y_m$ , their true score variables, and their error variables, respectively.

The assumption (a<sub>1</sub>) to (a<sub>3</sub>) specify in different ways the assumption that two tests  $Y_i$  and  $Y_j$  measure the same attribute. Such an assumption is crucial for inferring the degree of reliability from the discrepancy between two measurements of the same attribute of the same person. Perfect identity or ' $\tau$ -equivalence' of the two true score variables is assumed with (a<sub>1</sub>). With (a<sub>2</sub>) this assumption is relaxed: the two true score variables may differ by an additive constant. Two balances, for instance, will follow this assumption if one of them yields a weight that is always one pound larger than the weight indicated by the other balance, irrespective of the object to be weighed. According to Assumption (a<sub>3</sub>), the two tests measure the same attribute in the sense that their true score variables are linear functions of each other.

The other two assumptions deal with properties of the measurement errors. With (b) one assumes measurement errors pertaining to different test score variables to be uncorrelated. In (c) 'equal error variances are assumed,' i.e., these tests are assumed to measure equally well.

4.1 Parallel Tests

4.1.1 Definition. The most simple and convenient set of assumptions is the model of 'parallel tests.' Two tests  $Y_i$  and  $Y_j$  are defined to be parallel if they are  $\tau$ -equivalent, if their error variables are uncorrelated, and if they have identical error variances. Note that Assumption (a<sub>1</sub>) implies that there is a uniquely defined latent variable being identical to each of the true score variables. Hence, one may drop the

**Table 4**  
The Model of Parallel Tests

Definition	Assumptions (a <sub>1</sub> ), (b) and (c) of Table 3
Identification	$E(\eta) = E(Y_i)$ $Var(\eta) = Cov(Y_i, Y_j), i \neq j$ $Var(\varepsilon_i) = Var(Y_i) - Cov(Y_i, Y_j), i \neq j$ $Rel(Y_i) = Corr(Y_i, Y_j), i \neq j$
Testability	
in the total population	$E(Y_i) = \mu$ $Var(Y_i) = \sigma_Y^2$ $Cov(Y_i, Y_j) = \sigma_\eta^2$
within each subpopulation $s$	$E^s(Y_i) = \mu_s$

*Note:* The indices  $i$  and  $j$  refer to tests and the superscripts  $s$  to a subpopulation. The equations are true for each test  $Y_i$  of a set of parallel tests  $Y_1, \dots, Y_m$  or for each pair of two such tests, their true score variables, and their error variables, respectively.

index  $i$  and denote this latent variable by  $\eta$ . The assumption of  $\tau$ -equivalence may equivalently be written  $Y_i = \eta + \varepsilon_i$ , where  $\varepsilon_i := Y_i - E(Y_i|U)$ .

**4.1.2 Identification.** For parallel tests the theoretical parameters may be computed from the parameters characterizing the distribution of at least two test score variables, i.e., the theoretical parameters are identified in this model if  $m \geq 2$ . According to Table 4 the expected value of  $\eta$  is equal to the expected value of each of the tests, whereas the variance of  $\eta$  can be computed from the covariance of two different tests. The variance  $Var(\varepsilon_i)$  of the measurement error variables may be computed by the difference  $Var(Y_i) - Cov(Y_i, Y_j), i \neq j$ . Finally, the reliability  $Rel(Y_i)$  is equal to the correlation  $Corr(Y_i, Y_j)$  of two different test score variables.

**4.1.3 Testability.** The model of parallel tests implies several consequences that may be tested empirically. First, all parallel tests  $Y_i$  have equal expectations  $E(Y_i)$ , equal variances  $Var(Y_i)$ , and equal covariances  $Cov(Y_i, Y_j)$  in the total population. Second, parallel tests also have equal expectations within each subpopulation (see Table 4).

Note that these hypotheses may be tested separately and/or simultaneously as a single multidimensional hypothesis in the framework of 'simultaneous equation models' via AMOS (Arbuckle 1997), EQS (Bentler 1995), LISREL 8 (Jöreskog and Sörbom 1998), MPLUS (Muthén and Muthén 1998), MX (Neale 1997), RAMONA (Browne and Mels 1998), SEPATH (Steiger 1995), and others. Such a simultaneous test may even include the hypotheses about the parameters in several subpopulations (see Table 4). What is not implied by the assumptions of parallel tests is the equality of the variances and the covariances of the test score variables in subpopulations.

For parallel tests  $Y_1 + \dots + Y_m$  as defined in Table 4, the reliability of the sum score  $S := Y_1 + \dots + Y_m$  may

be computed by the 'Spearman-Brown formula for lengthened tests:'

$$Rel(S) = Rel(S/m) = \frac{m \cdot Rel(Y_i)}{1 + (m-1) \cdot Rel(Y_i)}$$

Using this formula the reliability of an aggregated measurement consisting of the sum (or average) of  $m$  parallel measurements of the same unit can be computed. For  $m = 2$ , each with  $Rel(Y_i) = 0.80$ , for instance

$$Rel(S) = 2 \cdot 0.80 / 1 + (2-1) \cdot 0.80 \approx 0.89$$

The Spearman-Brown formula may also be used to answer the opposite question. Suppose there is a test being the sum of  $m$  parallel tests and this test has reliability  $Rel(S)$ . What would be the reliability  $Rel(Y_i)$  of the  $m$  parallel tests? For example, if  $m = 2$ , what would be the reliability of a test half?

## 4.2 Essentially $\tau$ -equivalent Tests

**4.2.1 Definition.** The model of essentially  $\tau$ -equivalent tests is less restrictive than the model of parallel tests. Two tests  $Y_i$  and  $Y_j$  are defined to be 'essentially  $\tau$ -equivalent' if their true score variables differ only by an additive constant (Assumption a<sub>2</sub> in Table 3) and if their error variables are uncorrelated (Assumption b in Table 3). Assumption (a<sub>2</sub>) implies that there is a latent variable  $\eta$  that is a translation of each of the true score variables, i.e.,

$$\eta = \tau_i + \lambda_i, \lambda_i \in \mathbb{R}, \text{ such that } Y_i = \eta + \lambda_i + \varepsilon_i$$

where  $\varepsilon_i := Y_i - E(Y_i|U)$  and  $\lambda_i \in \mathbb{R}$

Also note that the latent variable  $\eta$  is uniquely defined up to a translation. Hence, it is necessary to fix the scale of the latent variable  $\eta$ . This can be done by fixing one of the coefficients  $\lambda_i$  (e.g.,  $\lambda_1 = 0$ ) or by fixing the expected value of  $\eta$  [e.g.,  $E(\eta) = 0$ ].

**Table 5**  
The Model of Essentially  $\tau$ -Equivalent Tests

Definition	Assumptions (a <sub>2</sub> ) and (b) of Table 3
Fixing the scale of $\eta$	$E(\eta) = 0$
Identification	$Var(\eta) = Cov(Y_i, Y_j), i \neq j$ $Var(\varepsilon_i) = Var(Y_i) - Cov(Y_i, Y_j), i \neq j$ $Rel(Y_i) = Cov(Y_i, Y_j) / Var(Y_i), i \neq j$

Table 5 summarizes the assumptions defining the model and the consequences for identification and testability. In this model, the reliability cannot be identified any more by the correlation between two tests. Instead the reliability is identified by

$$Rel(Y_i) = Cov(Y_i, Y_j) / Var(Y_i), \quad i \neq j.$$

Furthermore, the expected values of different tests are not identical any more within each subpopulation. Instead, the differences between the expected values  $E^{(s)}(Y_i) - E^{(s)}(Y_j)$  of two essentially  $\tau$ -equivalent tests  $Y_i$  and  $Y_j$  are the same in each and every subpopulation. All other properties are the same as in the model of parallel tests. Again, all these hypotheses may be tested via structural equation modeling.

For essentially  $\tau$ -equivalent tests  $Y_1, \dots, Y_m$ , the reliability of the sum score  $S = Y_1 + \dots + Y_m$  may be computed by the 'Cronbach's coefficient  $\alpha$ ':

$$\alpha = \frac{m}{m-1} \left( 1 - \frac{\sum_{i=1}^m Var(Y_i)}{Var(S)} \right). \quad (2)$$

This coefficient is a lower bound for the reliability of  $S$  if only uncorrelated errors are assumed.

### 4.3 $\tau$ -Congeneric Tests

**4.3.1 Definition.** The model of  $\tau$ -congeneric tests is defined by the Assumptions (a<sub>3</sub>) and (b) in Table 3.

Hence, two tests  $Y_i$  and  $Y_j$  are called  $\tau$ -congeneric if their true score variables are positive linear functions of each other and if their error variables are uncorrelated. Assumption a<sub>3</sub> implies that there is a latent variable  $\eta$  such that each true score variable is a positive linear function of it, i.e.,

$$\tau_i = \lambda_{i0} + \lambda_{i1}\eta, \quad \lambda_{i0}, \lambda_{i1} \in \mathbb{R}, \lambda_{i1} > 0,$$

or equivalently:

$$Y_i = \lambda_{i0} + \lambda_{i1}\eta + \varepsilon_i,$$

where  $\varepsilon_i = Y_i - E(Y_i|U)$ .

The latent variable  $\eta$  is uniquely defined up to positive linear functions. Hence, in this model, too, it is necessary to fix the scale of  $\eta$ . This can be done by fixing a pair of the coefficients (e.g.,  $\lambda_{i0} = 0$  and  $\lambda_{i1} = 1$ ) or by fixing the expected value and the variance of  $\eta$  [e.g.,  $E(\eta) = 0$  and  $Var(\eta) = 1$ ].

Table 6 summarizes the assumptions defining the model and the consequences for identification and testability assuming  $E(\eta) = 0$  and  $Var(\eta) = 1$ . Other ways of fixing the scale of  $\eta$  would imply different formula. As can be seen from the formula in Table 6, the model of  $\tau$ -congeneric variables and all its parameters are identified if there are at least three different tests for which Assumptions a<sub>3</sub> and b hold. The covariance structure in the total population implied by the model may be tested empirically if there are at least four test score variables. Only in this case the model has fewer theoretical parameters determining the covariance matrix of the test score variables than there elements in this covariance matrix. The implications for the mean structure are testable already for three test score variables provided the means of the test score variables are available in at least four subpopulations.

### 4.4 Other Models of CTT

The models treated previously are not the only ones that can be used to determine the theoretical para-

**Table 6**  
The model of  $\tau$ -Congeneric Tests

Definition	Assumptions (a <sub>3</sub> ) and (b) of Table 3
Fixing the scale of $\eta$	$E(\eta) = 0$ and $Var(\eta) = 1$
Identification	$\lambda_{i1} = \sqrt{\frac{Cov(Y_i, Y_j) Cov(Y_i, Y_k)}{Cov(Y_j, Y_k)}}, i \neq j, i \neq k, j \neq k$ {eqn6.eps} $Var(\varepsilon_i) = Var(Y_i) - \lambda_{i1}^2$ $Rel(Y_i) = \lambda_{i1}^2 / Var(Y_i)$
Testability in the total population	$\frac{Cov(Y_i, Y_k)}{Cov(Y_j, Y_k)} = \frac{Cov(Y_i, Y_l)}{Cov(Y_j, Y_l)}, i \neq k, i \neq l, j \neq k, j \neq l$ {eqn7.eps}
between subpopulations	$\frac{E^{(1)}(Y_i) - E^{(2)}(Y_i)}{E^{(1)}(Y_j) - E^{(2)}(Y_j)} = \frac{E^{(3)}(Y_i) - E^{(4)}(Y_i)}{E^{(3)}(Y_j) - E^{(4)}(Y_j)}$ {eqn8.eps}

Note: The indices  $i$  and  $j$  refer to tests and the superscripts to one of four subpopulations.

meters of CTT such as reliability, true score variance, and error variance. In fact, the models dealt with are limited to unidimensional models. However, true score variables may also be decomposed into several latent variables. 'Confirmatory factor analysis' provides a powerful methodology to construct, estimate, and test models with multidimensional decompositions of true score variables. Note, however, that not each factor model is based on CTT. For instance, there are one-factor models that are not models of  $\tau$ -congeneric variables in terms of CTT. A model with one common factor and several specific but uncorrelated factors is a counter example. The common factor is not necessarily a linear function of the true score variables and the specific factors are not necessarily the measurement error variables as defined in CTT.

#### 4.5 Some Practical Issues

Once a measurement or test score has been obtained for a specific individual, one might want to know how dependable that individual measurement is. If the reliability of the measurement is known and if one assumes a normal distribution of the measurement errors which is homogeneous for all individuals, the 95 percent-confidence interval for the true score of that individual with respect to the measurement  $Y_i$  can be computed by:

$$Y_i \pm 1.96 \cdot \sqrt{\text{Var}(Y_i) \cdot (1 - \text{Rel}(Y_i))}. \quad (3)$$

Another result deals with the correlation between two true score variables. If the reliabilities for two test score variables  $Y_1$  and  $Y_2$  are known, and assuming uncorrelated measurement errors, one may compute

$$\text{Corr}(\tau_1, \tau_2) = \frac{\text{Corr}(Y_1, Y_2)}{\sqrt{\text{Rel}(Y_1)} \cdot \sqrt{\text{Rel}(Y_2)}}. \quad (4)$$

This eqn. is known as the 'correction for attenuation.'

Another important issue deals with the 'reliability of a difference variable,' for example, a difference between a pretest  $Y_1$  and a posttest  $Y_2$ . Assuming equal true score and error variances between pre- and posttest implies identical reliabilities, i.e.,  $\text{Rel}(Y_1) = \text{Rel}(Y_2) = \text{Rel}(Y)$ . If additionally uncorrelated measurement errors are assumed, the reliability  $\text{Rel}(Y_1 - Y_2) := \text{Var}(E(Y_1 - Y_2|U)) / \text{Var}(Y_1 - Y_2)$  of the difference  $Y_1 - Y_2$  may be computed by:

$$\text{Rel}(Y_1 - Y_2) = \frac{\text{Rel}(Y) - \text{Corr}(Y_1, Y_2)}{1 - \text{Corr}(Y_1, Y_2)} \quad (5)$$

According to this formula, the reliability of a difference between pre- and posttest is always smaller than the reliability of the pre- and posttest, provided the assumptions mentioned above hold. In the extreme

case in which there is no differential change, i.e.,  $\tau_2 = \tau_1 + \text{constant}$ , the reliability coefficient  $\text{Rel}(Y_1 - Y_2)$  will be zero. Obviously, this does not mean that the change is not dependable. It only means that there is no variance in the change, since each individual changes by the same amount. This phenomenon has led to much confusion about the usefulness of measuring change (e.g., Cronbach and Furby 1970, Harris 1963, Rogosa 1995). Most of these problems are now solved by structural equation modeling (allowing the include latent change variables such as in growth curve models (e.g., McArdle and Epstein 1987, Willet and Sayer 1996) or, more directly, in true change models (Steyer, Eid, and Schwenkmezger 1997, Steyer, Partchev, and Shanahan, 2000). Models of this kind are no longer hampered by reliability problems and allow the explanation of inter-individual differences in intraindividual change.

#### 5. Discussion

It should be noted that CTT refers to the population level, i.e., to the random experiment of sampling a single observational unit and assessing some of its behavior (see Table 1). CTT does not refer to sampling models that consist of repeating this random experiment many times. Hence, no questions of statistical estimation and hypothesis testing are dealt with. Of course, the population models of CTT have to be supplemented by sampling models when it comes to applying statistical analyses, e.g., via structural equation modeling.

Aside from this more technical aspect, what are the limitations of CTT? First, CTT and its models are not really adequate for modeling answers to individual items in a questionnaire. This purpose is more adequately met by models of item response theory (IRT) which specify how the probability of answering in a specific category of an item depends on the attribute to be measured, i.e., on the value of a latent variable.

A second limitation of CTT is the exclusive focus on measurement errors. 'Generalizability theory' presented by Cronbach et al. (1972) (see also Shavelson and Webb 1991) generalized CTT to include other factors determining test scores.

Inspired by Generalizability Theory Tack (1980), Steyer et al. (1989) presented a generalization of CTT, called 'Latent State-Trait Theory,' which explicitly takes into account the situation factor, introduced formal definitions of states and traits, and presented models allowing to disentangle person, as well as situation and/or interaction effects and from measurement error. More recent presentations are Steyer et al. (1992) as well as Steyer, et al. (1999). Eid (1995, 1996) extended this approach to the normal ogive model for analyses on the item level.

The parameters of CTT are often said to be 'population dependent,' i.e., meaningful only with respect to a given population. This is true for the

variance of the true score variable and the reliability coefficient. The reliability (coefficient) of an intelligence test is different in the population of students than the general population. This is a simple consequence of the restriction of the (true score) variance of intelligence in the population of the students. However, such a restriction neither exists for the true score estimates neither of individual persons nor for the item parameters  $\lambda_i$  of the model of 'essentially  $\tau$ -equivalent tests,' for instance. Proponents of IRT models have often forwarded the population dependence critique. They contrast it with the 'population independence' of the person and the item parameters of IRT models. However, 'population independence' also holds for the person and the item parameters of the model of essentially  $\tau$ -equivalent tests, for instance.

In applications of CTT it is often assumed that the error variances are the same for each individual, irrespective of the true score of that individual. This assumption may indeed be wrong in many applications. In IRT models no such assumption is made. However, it is possible to assume different error variances for different (categories of) persons in CTT models as well. In this case, the unconditional error variance and the reliability coefficient are not the best available information for inferring the dependability of individual true score estimates. In this case one should seek to obtain estimates of conditional measurement error variances for specific classes of persons. It is to be expected that persons with high true scores have a higher measurement error variance than those with medium true scores and that those with low true scores have a higher error variance again. (This would be due to 'floor and ceiling effects.'). Other patterns of the error variance depending on the size of the true score may occur as well.

Such phenomena do not mean that 'true scores' and 'error scores' would be correlated; only the 'error variances' would depend on the true scores. None of the properties listed in Table 2 would be violated. As mentioned before, the properties listed in Table 2 cannot be wrong in empirical applications. What could be wrong, however, is the true score interpretation of the latent variable in a concrete structural equation model. Misinterpretations of this sort can be most effectively prevented by empirical tests of the hypotheses listed in the testability sections of Tables 4 to 6.

The most challenging critique of many applications of CTT is that they are based on rather arbitrarily defined test score variables. If these test score variables are not well chosen any model based on them is also not well founded. Are there really good reasons to base models on sum scores across items in questionnaires? Why take the sum of the items as test score variables  $Y_i$  and not another way of aggregation such as a weighted sum, or a product, or a sum of logarithms? And why aggregate and not look at the items themselves?

There is no doubt that IRT models are more informative than CTT models if samples are big enough to allow their application, if the items obey the laws defining the models, and if detailed information about the items (and even about the categories of 'polytomous items,' such as in 'ratings scales') is sought. In most applications, the decision how to define the test score variables  $Y_i$  on which models of CTT are built is arbitrary, to some degree. It should be noted, however, that arbitrariness in the choice of the test score variables cannot be avoided altogether. Even if models are based on the item level, such as in IRT models, one may ask 'Why these items and not other ones?' Whether or not a good choice has been made will only prove in model tests and in validation studies. This is true for models of CTT as well as for models of alternative theories of psychometric tests.

*See also:* Dimensionality of Tests: Methodology; Factor Analysis and Latent Structure: IRT and Rasch Models; Generalizability Theory; Psychometrics; Reliability: Measurement; Test Theory: Applied Probabilistic Measurement Structures

#### Bibliography

- Arbuckle J L 1997 *Amos User's Guide: Version 3.6*. SPSS, Chicago, IL
- Bentler P M 1995 *EQS. Structural Equation Program Manual*. Multivariate Software, Encino
- Browne M W, Mels G 1998 Path analysis (RAMONA). In: SYSTAT 8.0 -Statistics. SPSS, Inc., Chicago
- Cronbach L J, Furby L 1970 How should we measure 'change'—or should we? *Psychological Bulletin* 74: 68–80
- Cronbach L J, Gleser G C, Nanda H, Rajaratnam N 1972 *The Dependability of Behavioral Measurements: Theory of Generalizability of Scores and Profiles*. Wiley, New York
- du Toit M, du Toit S 2001 *Interactive LISREL: User's Guide*. Scientific Software International, Chicago
- Eid M 1995 *Modelle der Messung von Personen in Situationen* [Models of measuring persons in situations]. Psychologie Verlags Union, Weinheim, Germany
- Eid M 1996 Longitudinal confirmatory factor analysis for polytomous item responses: model definition and model selection on the basis of stochastic measurement theory [online]. <http://www.ppm.ipn.uni-kiel.de/mpr/issue1/art4/eid.pdf>: 1999-7-5
- Fischer G H, Molenaar I W 1995 *Rasch Models*. Springer, New York
- Gulliksen H 1950 *Theory of Mental Tests*. Wiley, New York
- Jöreskog K G, Sörbom D 1998 *LISREL 8. Users Reference Guide*. Scientific Software, Chicago
- Lord F M, Novick M R 1968 *Statistical Theories of Mental Test Scores*. Addison Wesley, Reading, MA
- Muthén L, Muthén B 1998 *Mplus User's Guide*. Muthén & Muthén, Los Angeles
- Neale M C 1997 *MX: Statistical Modeling*, 4th edn. Department of Psychiatry, Richmond, VA
- Novick M R 1966 The axioms and principal results of classical test theory. *Journal of Mathematical Psychology* 3: 1–18
- Rogosa D 1995 Myths and methods: 'Myths about longitudinal

- research' plus supplemental questions. In: Gottman J M (ed.) *The Analysis of Change*. Lawrence Erlbaum, Mahwah, NJ
- Shavelson R J, Webb N M 1991 *Generalizability Theory. A Primer*. Sage, Newbury Park
- Steiger J H 1995 Structural equation modeling. In: STATISTICA 5—Statistics II. StatSoft Inc., Tulsa, OK
- Steyer R, Ferring D, Schmitt M J 1992 States and traits in psychological assessment. *European Journal of Psychological Assessment* 8: 79–98
- Steyer R, Majcen A-M, Schwenkmezger P, Buchner A 1989 A latent state-trait anxiety model and its application to determine consistency and specificity coefficients. *Anxiety Research* 1: 281–99
- Steyer R, Partchev I, Shanahan M 2000 Modeling true intra-individual change in structural equation models: the case of poverty and children's psychosocial adjustment. In: Little T D, Schnabel K U, Baumert J (eds.) *Modeling Longitudinal and Multiple-Group Data: Practical Issues, Applied Approaches, and Specific Examples*. Erlbaum, Hillsdale, NJ, pp. 109–26
- Steyer R, Schmitt M, Eid M 1999 Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*
- Tack W H 1980 Zur Theorie psychometrischer Verfahren: Formalisierung der Erfassung von Situationsabhängigkeit und Veränderung [On the theory of psychometric procedures: formalizing the assessment of situational dependency and change]. *Zeitschrift für Differentielle und Diagnostische Psychologie* 1: 87–106
- Thurstone L L 1931 *The Reliability and Validity of Tests*. Edwards Brothers, Ann Arbor
- Zimmerman D W 1975 Probability spaces, hilbert spaces, and the axioms of test theory. *Psychometrika* 40: 395–412
- Zimmerman D W 1976 Test theory with minimal assumptions. *Educational and Psychological Measurement* 36: 85–96
- Zimmerman D W, Williams R H 1977 The theory of test validity and correlated errors of measurement. *Journal of Mathematical Psychology* 16: 135–52

R. Steyer

## Classification and Typology (Archaeological Systematics)

Classification is the initial means through which we impose a degree of order on the enormously diverse remains of the human past. As such, it is probably the single most basic analytical procedure employed by the archaeologist. Excavation yields an enormous diversity of materials that are not self-labeling; they must be endowed with identity and meaning by the excavator or the analyst. This is done in the first instance through classification.

### 1. Classification and Typology

Archaeologists often use the terms classification and typology interchangeably, but in this article a distinction will be made. A classification is any set of formal categories into which a particular field of data

is partitioned, while a typology is a particular type of rigorous classification, in which a field of data is divided up into categories that are all defined according to the same set of criteria, and that are mutually exclusive. As will be shown later, most archaeological classifications of artifacts are typologies, while most classifications of cultures are not.

### 1.1 Archaeological Classification and Culture

The basic organizing concept for most prehistorians, as for most other anthropologists, is the concept of culture, but it is somewhat differently defined in the two cases. The cultural anthropologist conceives of the world as divided into a set of distinct peoples—tribes, nations, or ethnic groups—each of which has its own unique set of behavior patterns and beliefs, very often including its own language, which together constitute a culture. The prehistorian thinks of the ancient world as similarly partitioned, but the various long-vanished peoples can now be recognized only by the distinct kinds of artifact types they left behind. In place of forgotten languages and behavior patterns, every artifact type is treated as tantamount to a deliberate cultural expression—a culture trait. An archaeologically defined 'culture' is then a unique combination of artifact, house, and burial types, which are assumed, because of their cultural commonality, to be the remains left by a distinct, self-recognizing people. Those commonalities are recognized above all through processes of classification.

### 1.2 Kinds of Archaeological Classification

Obviously, any of the different kinds of material remains that archaeologists find can be classified, and there are in fact many different kinds of archaeological classifications and typologies. In the broadest sense, all of them fall into two categories, which may be called analytic and synthetic. Analytic classifications are classifications of one particular kind of object, in which all of the regularly recurring variants are recognized, defined, and named. The things most often classified are those that show a high degree of culturally patterned variability, including various kinds of stone tools and weapons; pottery; beads and other ornaments; house types; and grave types. Classifications of these things are usually typologies; that is, they partition the entire field of variability into a comprehensive set of mutually exclusive categories, because they are very commonly used for sorting and counting the objects found.

Artifact typologies can be made in a wide variety of ways, depending on what criteria of identity are considered important. This in turn will depend on the purpose for which the classification is made. Among the many kinds of artifact classifications it is possible to recognize purely morphological typologies, based on the overall form of objects; stylistic typologies,