

## CHAPTER SIX: CHARACTERISTICS OF A GOOD TEST

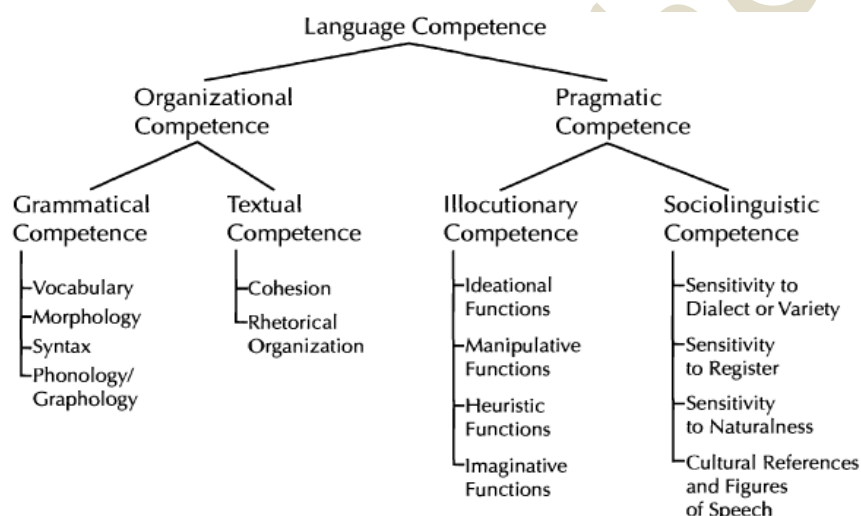
### 1. RELIABILITY

# The notion of *consistency of one's score with respect to one's average score over repeated administrations* is the central concept of reliability

# A test is reliable if it would give us the same result over and over again (assuming that what we are measuring is not changing)

# If one takes two measures of the same attribute, the two measures will not resemble each other exactly:

- (a) **meaningful variance** → those creating variance related to *the purposes of the test* (score change due to learning/forgetting) → predictable, systematic

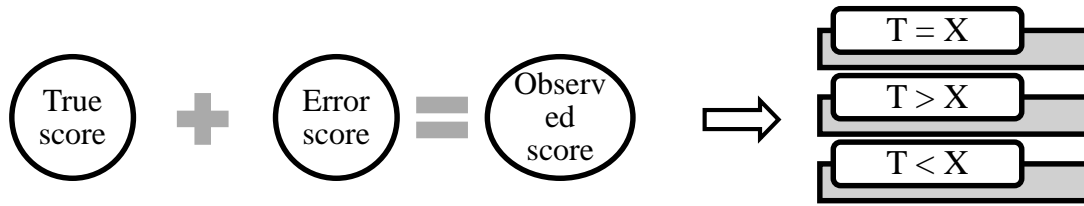


- (b) **error variance** → those generating variance due to *other extraneous sources* → unpredictable, unsystematic

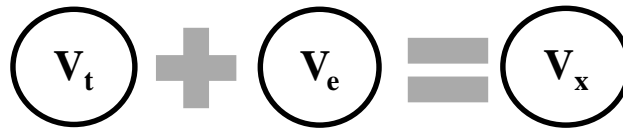
<b>Variance due to environment</b> <ul style="list-style-type: none"> <li>• ventilation</li> <li>• noise</li> <li>• lighting</li> <li>• weather</li> </ul>	<b>Variance due to administration procedures</b> <ul style="list-style-type: none"> <li>• directions</li> <li>• equipment</li> <li>• timing</li> </ul>	<b>Variance due to scoring procedures</b> <ul style="list-style-type: none"> <li>• errors in scoring</li> <li>• subjectivity</li> <li>• evaluator biases</li> </ul>
<b>Variance due to the test and test items</b> <ul style="list-style-type: none"> <li>• test booklet clarity</li> <li>• particular sample of items</li> <li>• number of items</li> </ul>	<b>Variance attributable to examinees</b> <ul style="list-style-type: none"> <li>• health</li> <li>• fatigue</li> <li>• motivation</li> </ul>	

### 2. CLASSICAL TRUE SCORE THEORY (CTS)

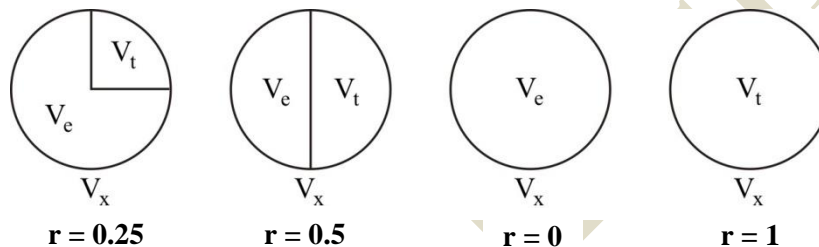
# An **observed score** comprises two factors or components: a **true score** and an **error score**. The relationship between the observed and true score can be illustrated as follows:



We don't speak of the reliability of a measure for 'an individual' – reliability is a characteristic of a measure that is taken 'across individuals':



**Reliability** → the ratio of the variance of true scores to the variance of observed scores:  $r = \frac{V_t}{V_x}$



### 3. STANDARD ERROR OF MEASUREMENT

The formula for calculating SEM is relatively simple:

$$SEM = S_x \sqrt{1 - r}$$

where

$S_x$  = the standard deviation of the test

$r$  = reliability of the test

**Example:** If the standard deviation of a test were 15 and its reliability were estimated as 0.84, then what would be standard error of measurement?

$$SEM = S_x \sqrt{1 - r} = 15 \times \sqrt{1 - 0.84} = 15 \times \sqrt{0.16} = 15 \times 0.4 = 6$$

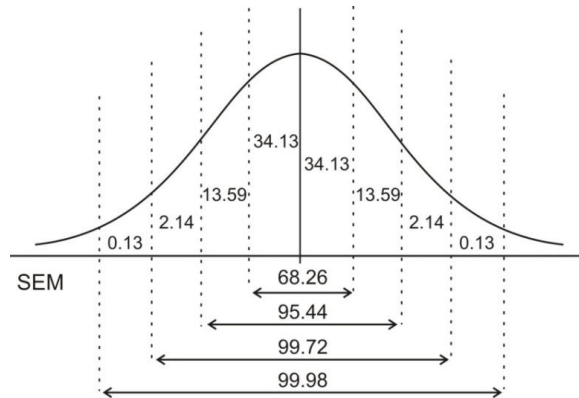
**Example:** When the test has no measurement error, then reliability would be -----.

$$SEM = S_x \sqrt{1 - r} \rightarrow 0 = S_x \times \sqrt{1 - r} \rightarrow 0 = \sqrt{1 - r} \rightarrow r = 1$$

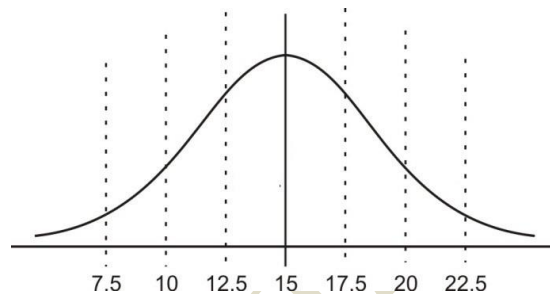
☛ **Note:** There is a negative relationship between standard error of measurement and reliability.

Conceptually → SEM provides a concrete estimate in test score values of the amount of unreliable score variation in a set of scores

Practically → SEM is used to determine a band around a student's observed score within which that student's true score would probably fall



**Example:** If the SEM of a set of scores is 2.5 we can be sure that a student's true score who obtained 15 would fluctuate 68% of times between -----



#### 4. APPROACHES TO ESTIMATING RELIABILITY

##### 4.1. Stability (Test-retest Method)

Administering a given test twice and then calculating the correlation between the two sets of scores, using Pearson product-moment correlation coefficient.

☛ **Note:** consistency of scores *over time* or *temporal stability*.

Drawbacks:

- Two administrations
- Human being abilities are most likely to change from one administration to another.
- **Practice effect:** after the first test, testees would naturally have a better performance on the second administration

##### 4.2. Equivalence (Parallel-forms Method)

Two similar forms of the same test are administered to a group of examinees just once → Using Pearson product moment formula, reliability is calculated.

Parallel forms:

He usually ..... tennis every day.

a) plays

b) play

c) playing

My brother often ..... a cup of tea every morning.

a) drinking

b) drinks

c) drink

Drawbacks:

- Constructing two parallel forms
- **Ordering effect** ≠ counterbalanced test design

	Time 1	Time 2
<b>Half I</b>	Form A	Form B
<b>Half II</b>	Form B	Form A

### 4.3. Internal Consistency

Uses information internal available in one administration of a single test. Assumptions:

- test scores are **unidimensional** → items of a test all measure the same, single ability // speededness
- test items are **homogeneous**
- the items or parts of a test are **locally independent**

#### 4.3.1. Split-half Methods

A single test is administered to a group of examinees → the test is divided into two equal halves

#### Spearman Brown estimate

The correlation (Pearson product-moment) between the two halves is an estimate of the test score reliability → *Spearman-Brown prophecy formula*:

$$r_{\text{total}} = \frac{2(r_{\text{half}})}{1 + (r_{\text{half}})}$$

**Example:** The reliability of half of a grammar test is calculated to be 0.35. By applying the Spearman Brown's prophecy formula, the total reliability would be -----.

$$r_{\text{total}} = \frac{2r_{\text{half}}}{1 + r_{\text{half}}} = \frac{2 \times .35}{1 + .35} = \frac{0.7}{1.35} = 0.51$$

Assumptions:

- The two halves are equivalent, i.e. they have equal means and variances
- The two halves are experimentally independent of each other

#### Guttman estimate

It does not assume equivalence of the halves; it does not require computing a correlation between them:

$$\alpha = 2 \left( 1 - \frac{S_{\text{odd}}^2 + S_{\text{even}}^2}{S_{\text{total}}^2} \right)$$

### 4.3.2. Item variance methods

#### KR-21 Method

Developed by Kuder and Richardson.

$$(KR - 21)r = \left(\frac{K}{K-1}\right) \left(1 - \frac{\bar{X}(K - \bar{X})}{KV}\right)$$

where: **K** = the number of the items in a test

Assumptions:

- items are of equal difficulty
- items are scored dichotomously (no weighting scheme)

☛ **Note:** called **rational equivalence**

☛ **Note:** the easiest and most frequently used method

☛ **Note:** an underestimate index of reliability

#### KR-20 Method

Avoids the problem of underestimating reliability:

$$KR - 20 = \left(\frac{K}{K-1}\right) \left(1 - \frac{\sum pq}{S_t^2}\right)$$

where: **pq (item variance)** =  $IF(1 - IF)$

Assumption:

- items are scored correct/incorrect (i.e., dichotomously or binary)
- ☛ **Note:** most accurate and flexible internal consistency method

#### Cronbach alpha method

Used with weighted items where examinees may receive partial credit:

$$\alpha = \left(\frac{K}{K-1}\right) \left(1 - \frac{\sum S_i^2}{S_t^2}\right)$$

where: **S<sub>i</sub><sup>2</sup>** = item variances for each individual item

	Assumption		Effect if assumption is violated	
	Equivalence	Independence	Equivalence	Independence
<b>Spearman-Brown</b>	Yes	Yes	Underestimate	Overestimate
<b>Guttman</b>	-	Yes	-	Overestimate
<b>Kuder-Richardson</b>	Yes	Yes	Underestimate	Overestimate

## 5. FACTORS INFLUENCING RELIABILITY

### 5.1. The Effect of Testees

- Psychological and physiological conditions
- Testees' Homogeneity
- Guessing: Educated guess vs. wild guess
- Test-wiseness → a test taker's capacity to utilize the characteristics and formats of the test and the test taking situation to guess the correct answer

### 5.2. The Effect of Test Factors

- Homogeneity of the items
- The speed with which the test is performed
- Ambiguity of instructions and items
- Discriminability
- Number of items

$$r_k = \frac{kr_1}{1 + (k-1)r_1}$$

$r_k$  = the test when adjusted to k times its original length

$r_1$  = the observed reliability of the test with its present length

$$k = \frac{\text{تعداد ثانویه سوالات}}{\text{تعداد اولیه سوالات}}$$

### 5.3. The Effect of Administration Factor

- The influence of the environment
- Quality and test timing

### 5.4. The Influence of Scoring Factors

The concern is **rater reliability** in case of subjective items:

- **Intra-rater reliability** → fluctuations of a *single scorer in scoring items twice* → unclear scoring criteria, fatigue, bias toward particular good and bad students, or simple carelessness
- **Inter-rater reliability** → fluctuations of *different scorers scoring a single test* → lack of adherence to scoring criteria, inexperience, inattention, or even preconceived biases

To avoid the effect of scoring:

- Provide a detailed scoring key
- Identify candidates by number, not name
- Train scorers
- Employ multiple, independent scoring
- Agree acceptable responses and appropriate scores at outset of scoring

## 6. VALIDITY

# Reliability: How much of an individual's test performance is due to measurement error, or to factors other than the language ability we want to measure? → minimizing the effects of these factors

# Validity: How much of an individual's test performance is due to the language ability we want to measure? → maximizing the effects of these abilities

**Validity** → the extent to which a test measures what it is supposed to measure.

### 6.1. Content Validity

# Degree of correspondence between the test content and the content of the materials to be tested:

- Subject matter = the topics
- Instructional objectives = degree of learning that students are supposed to achieve

# NRT → the extent to which a test contains a representative sample of the larger universe it is supposed to represent

# Main issue is *sampling* → *appropriateness* of the test sample

# It provides subjective information → to reduce subjectivity:

- Have the test reviewed by more than one expert
- Transfer the detailed definition onto a table of specification

### 6.2. Face Validity

The way the test looks to the examinees, test administrators, educators, and the like → this is not validity in the technical sense

- a well-constructed, expected format with familiar tasks,
- a test that is clearly doable within the allotted time limit,
- items that are clear and uncomplicated,
- directions that are crystal clear,
- task that relate to their course work (content validity), and
- a difficulty level that presents a reasonable challenge,
- no surprises in the test.

### 6.3. Criterion-related validity

Investigates the correspondence between the scores obtained from the newly-developed test and the scores obtained from some independent outside criteria → Pearson product-moment correlation

- **Concurrent validity:** a particular trait is administered *concurrently* with another well-known test
- **Predictive validity:** the two tests are administered with *some time interval*.

☛ **Note:** known as **empirical** or **statistical validity**

☛ **Note:** content of the criterion measure must be on the same domain as that of the new test

### 6.4. Construct Validity

# A test has construct validity to the extent to which the psychological reality of a trait or construct (like

language proficiency) can be established

# If a test has construct validity, it is capable of measuring certain specific characteristics in accordance with a theory of language behavior and learning

# The major problem with psychological constructs is that testers cannot take a construct out of a student's brain and show that a test is in fact measuring it. Experiment:

multitrait-multimethod studies

factor analytic techniques

structural equation modeling

think-aloud protocol

differential-groups studies

intervention studies

## 7. FACTORS INFLUENCING VALIDITY

- Directions
- Difficulty levels of the test
- Sample truncation
- Structure of items
- Arrangement of items and correct responses

## 8. PRACTICALITY

Have to do with physically putting tests into place in a program

- Ease of Test Administration
- Ease of Test Scoring
- Ease of Interpretation and Application
- Ease of Test Construction
- The Cost Issue

## 9. RELIABILITY VS. VALIDITY

# Reliability is a mathematical concept = it is score-dependent

# Validity is a relative term = it depends on the purpose of the test = it is test-dependent

- If a test is reliable, it may or may not be valid.
- If a test demonstrates a certain degree of validity, it is to some extent reliable.

## 9. EXTRA POINTS

**Coaching effect** → the effect on test scores of 'teaching to the test'. Coaching can be defined as short term instruction in test *wiseness* and in answering questions similar to those appearing on the target examination

**Test comprise effect** → the acquisition of prior knowledge of test content