



Reliability of Measurement

Mohammed TA, Omar Ph.D. PT

**Rehabilitation Health Science Department
College of Applied Medical Sciences-KSU**

Mommarar@ksu.edu.sa

0542115404



Objectives

Defines reliability and distinguish among the various types.

Explores ways of establishing reliability and how it can be reported using descriptive and statistical methods.

Outlines

What is reliability?

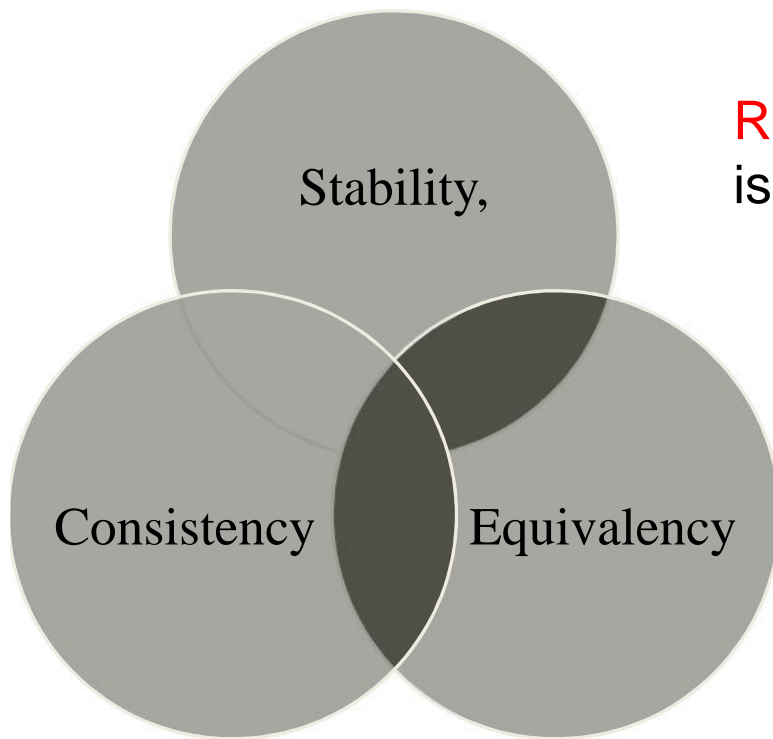
Describing types of reliability

- ❖ Test-retest reliability
- ❖ Internal consistency
- ❖ Parallel form reliability
- ❖ Split half reliability
- ❖ Intrarater reliability
- ❖ Interrater-reliability

How are studies of reliability analyzed?

- ❖ Percentage agreement and kappa
- ❖ Coefficients
- ❖ Intra-class correlation
- ❖ Bland and Altman method
- ❖ Internal consistency
- ❖ Standard error of the measurement

Reliability

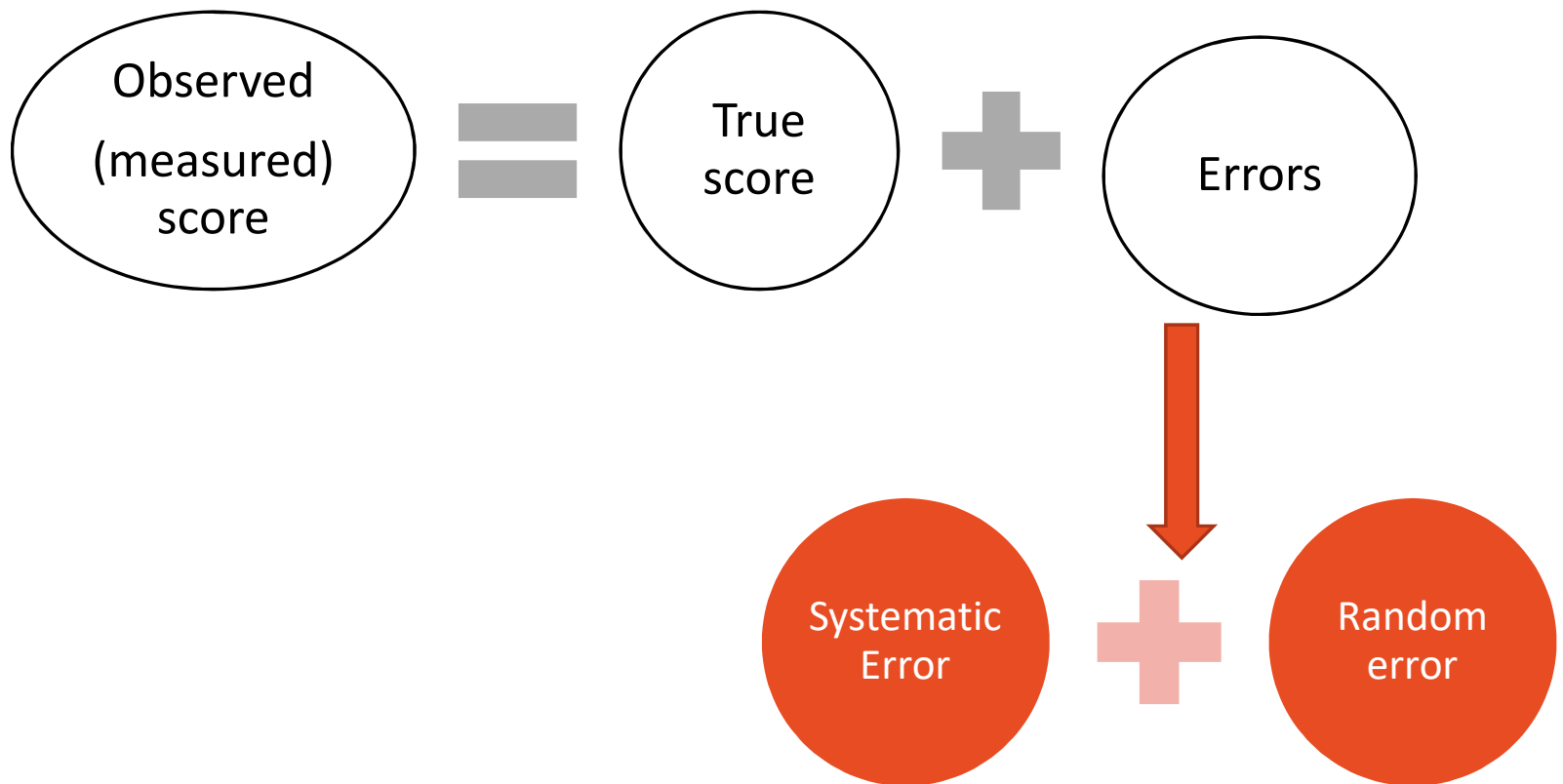


RELIABILITY is the degree to which a measure is free from **ERROR** & **CONSISTENT**.

RELIABILITY is not an all-or-none phenomenon,

Measurement Error

The term '**ERROR**' does not necessarily mean that something is being done **INCORRECTLY**; rather it refers to the variability in the technique of the person operating the instrument or making the measurement.



Measurement Error

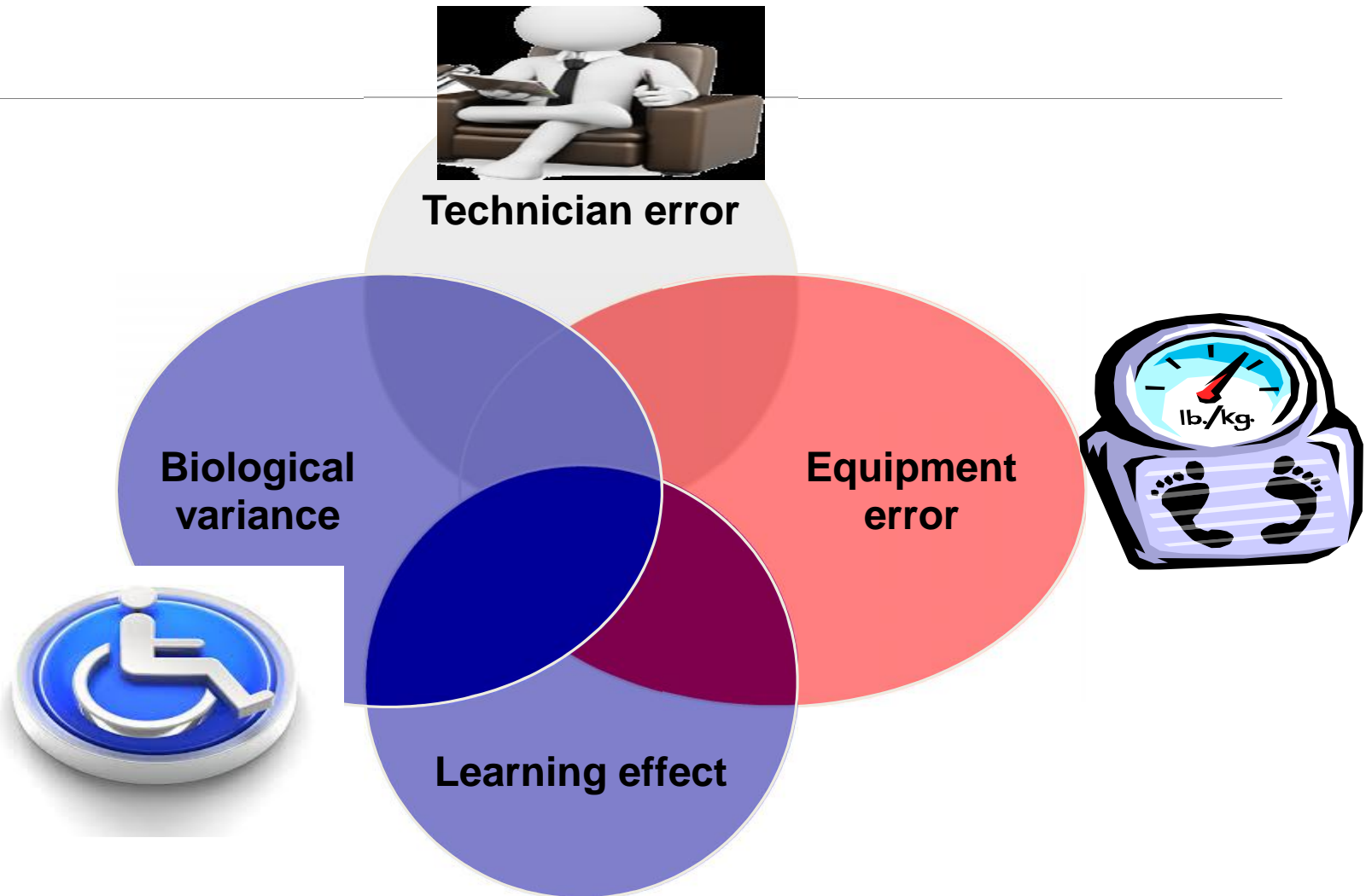
Random errors

- Are due to chance and unpredictable, thus they are the basic concern of reliability.

Systematic errors

- Systematic errors are predictable errors, occurring in one direction only, constant and biased

Sources of Measurement Errors



Sources of Error

Conditions of Test Administration and Construction

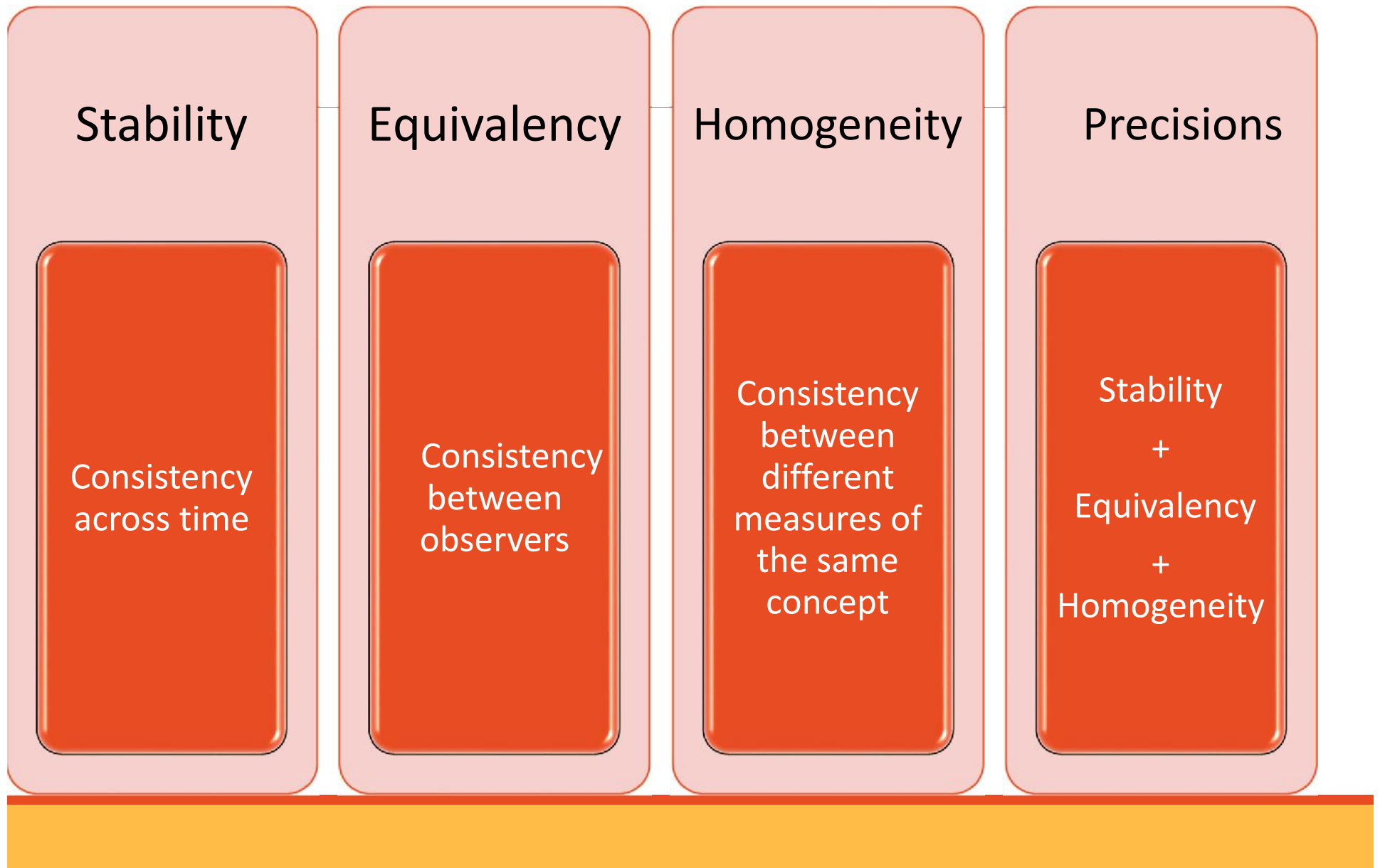
- Changes in time limits
- Changes in directions
- Different scoring procedures
- Interrupted testing session
- Qualities of test administrator
- Time test is taken
- Sampling of items
- unclearness in wording of items/questions
- Climate of test situation
(heating, light, ventilation, etc.)
- Differences in observers

Sources of Error

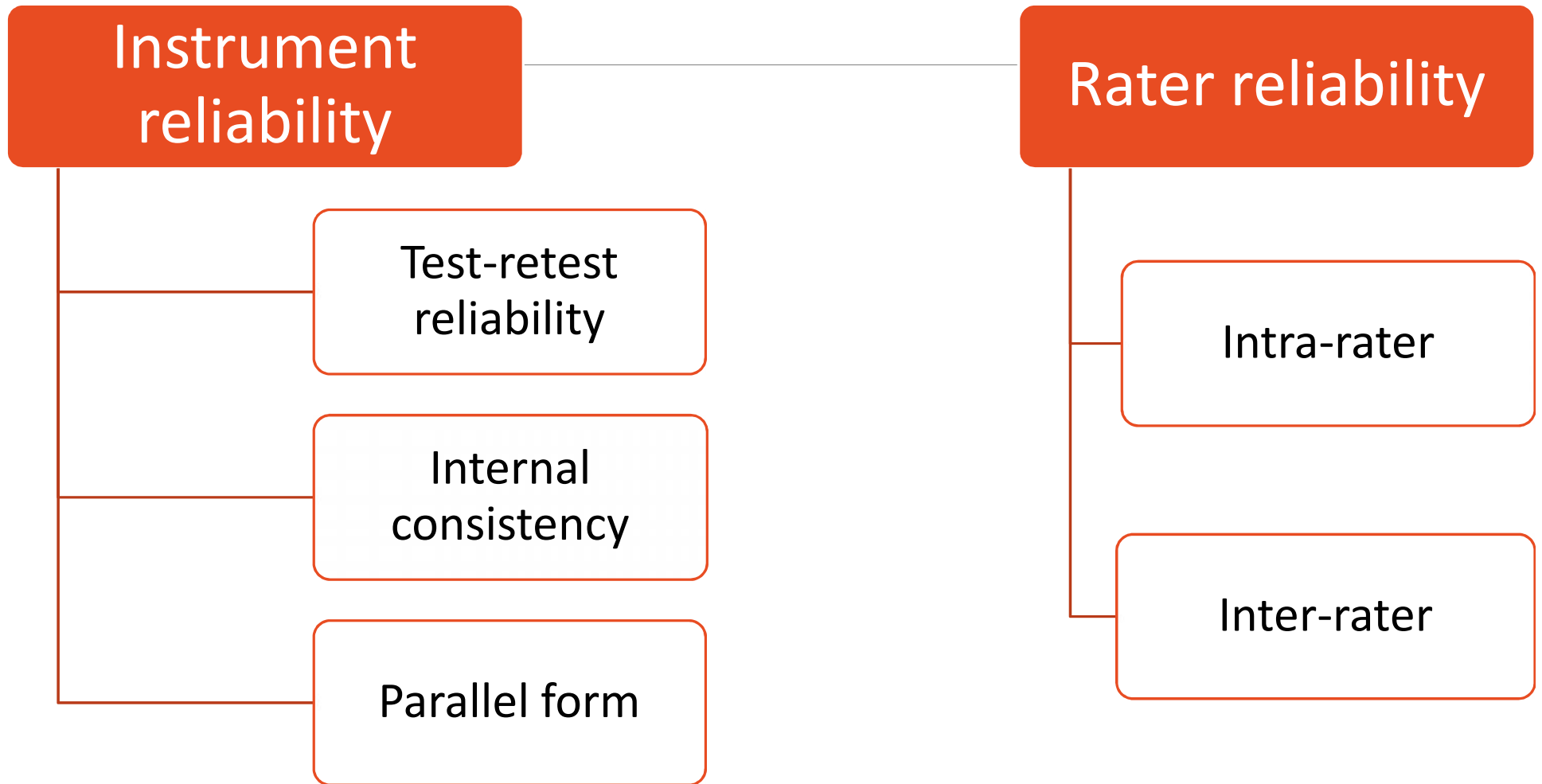
Conditions of the Person Taking the Test

- Reaction to specific items
- Health
- Motivation
- Mood
- Fatigue
- Luck
- Memory and/or attention fluctuations
- Attitudes
- Test-taking skills (test-wiseness)
- Ability to understand instructions
- Anxiety

Four Aspects of Reliability

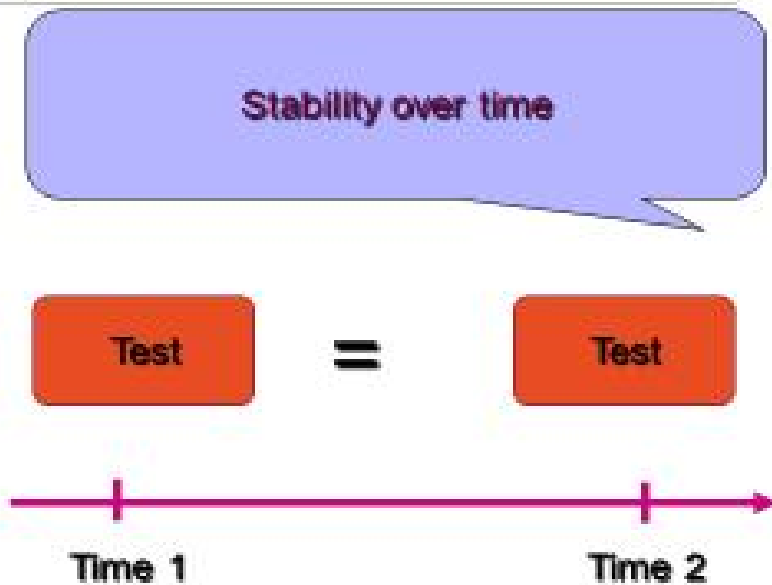


Types of Reliability



Test-Retest Reliability

Same raters/observers
+
Same groups/individual,
+
Used same measurement
+
At two different times.



- ❖ Consistency of patient/client results over time.
- ❖ To monitor changes following treatment.
- ❖ A single examiner can replicate the results

Test-Retest Reliability

The time lapse (2days-4weeks) depends on the type of test and behavior being evaluated.

A shorter period may be associated with

Patient recall of answers, increase skills of practice

Fatigue. changes in motivation,

Too long a period may be associated with Actual change in health.

Test-retest reliability is usually evaluated during questionnaire development.



Test-Retest Reliability


The interclass correlation coefficient (ICC) is the most frequently used method for estimates of test–retest reliability.

- ❑ For group comparisons, (ICC = 0.7)
- ❑ For individuals comparisons, (ICC = 0.9)

Internal Consistency Reliability

Internal consistency relates to the homogeneity of questions in the same domain and their ability to measure the same construct. It is most commonly associated with paper and pencil test

Internal consistency is the extent that certain items in a scale “hang together” and measure the same things.



Internal Consistency Reliability

The internal consistency reliability is important for four reasons

Consistency of results
across items within a test

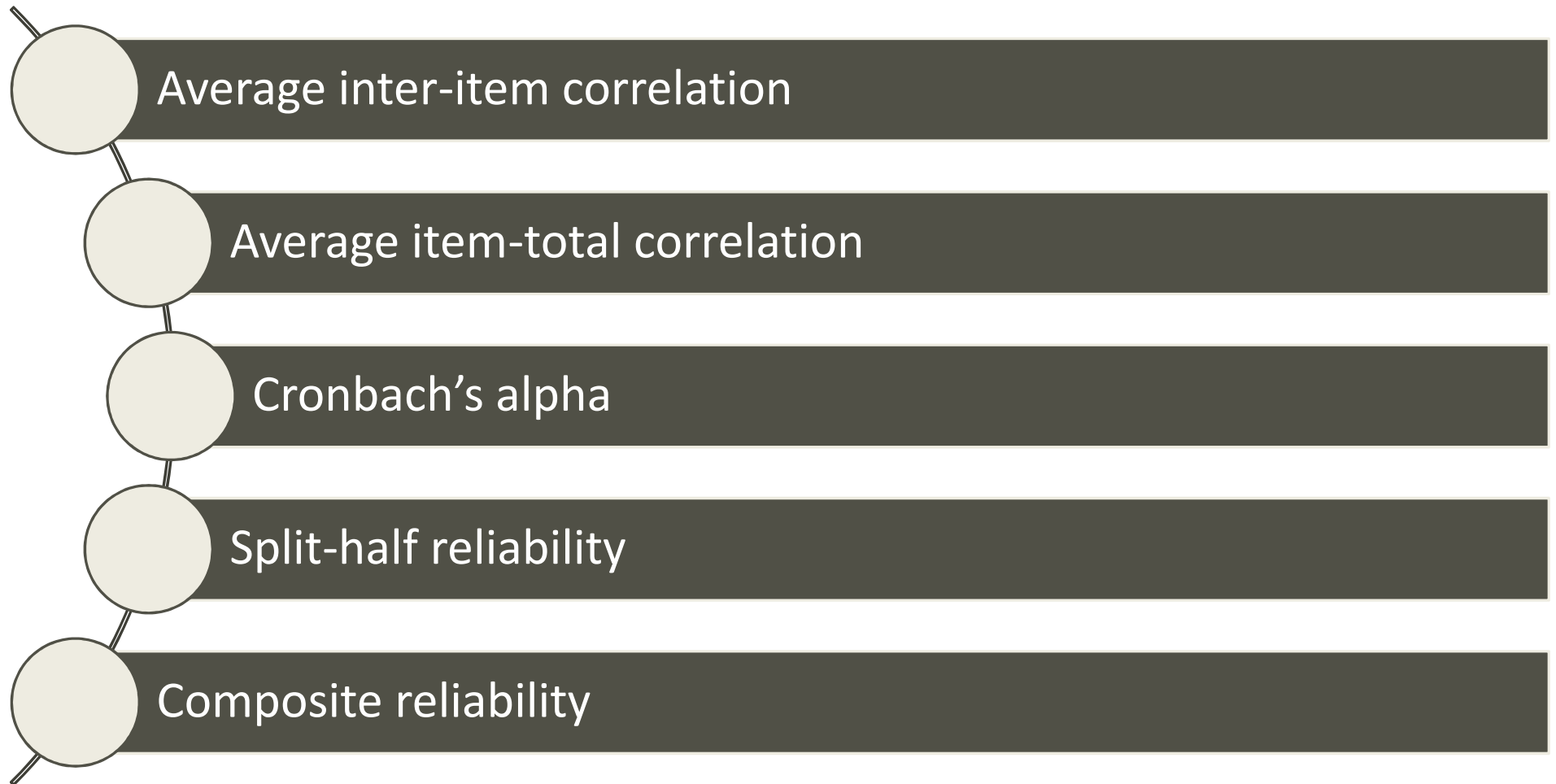
Assess the homogeneity
of items

Used to calculate the
standard error of
measures (SEM)

Internal consistency is frequently evaluated with Cronbach's alpha (α), which is generally deemed acceptable at values of 0.7-0.9.

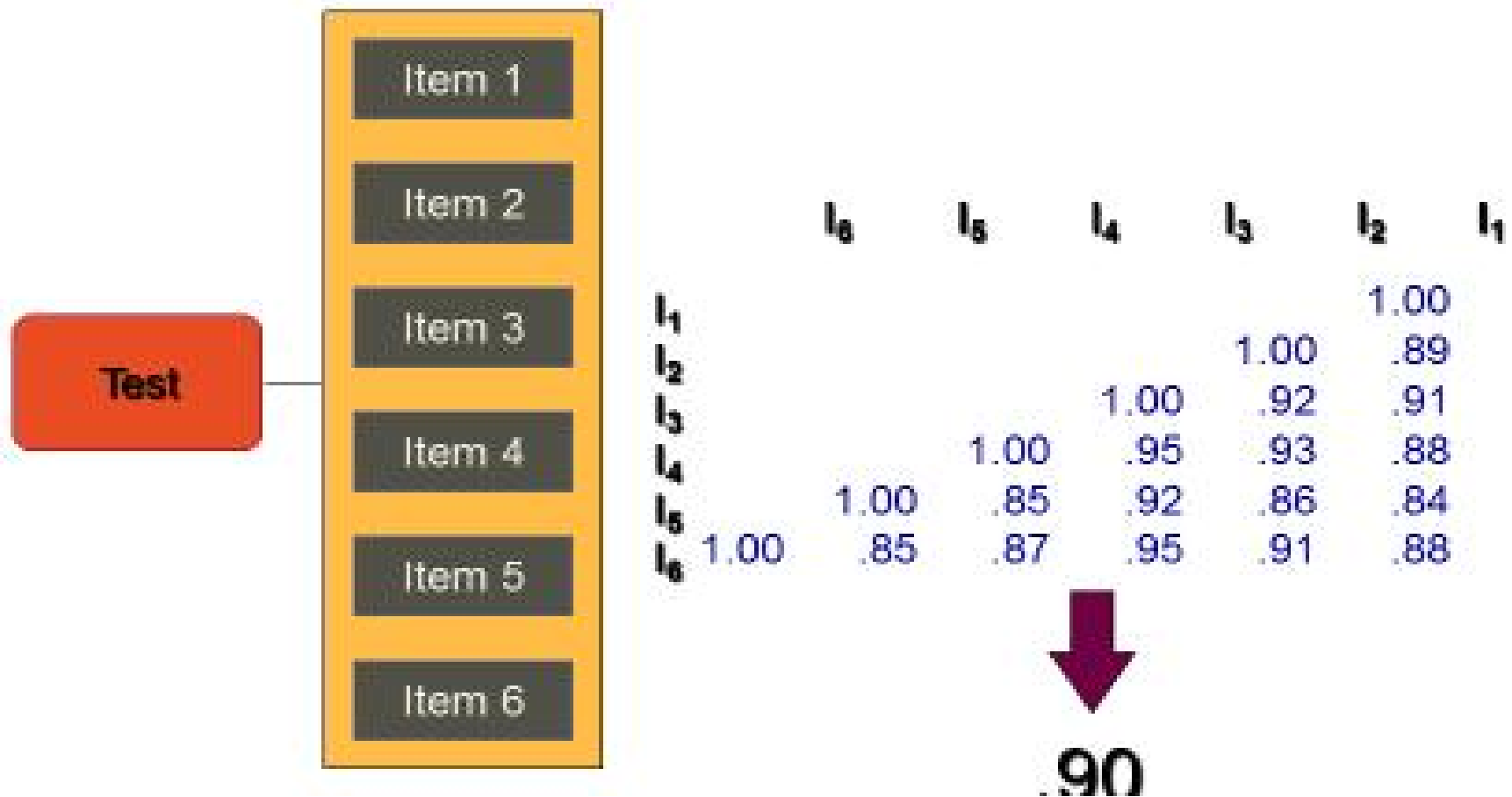
- ❖ The higher (α), the test is homogenous construct,
- ❖ The lower (α), the test has heterogenic factors

Five ways to calculate internal consistency



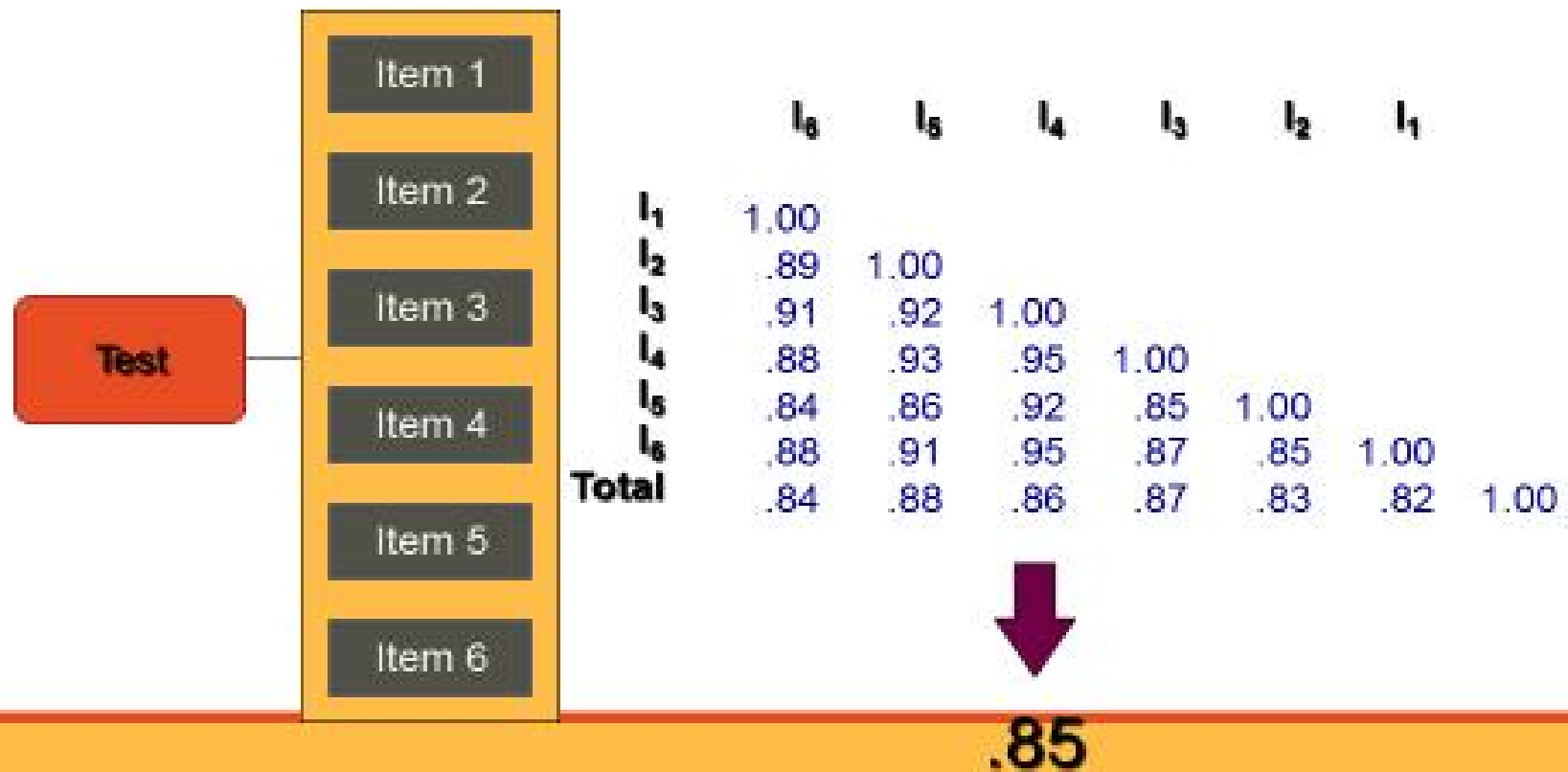
Internal Consistency Reliability

Average inter-item correlation



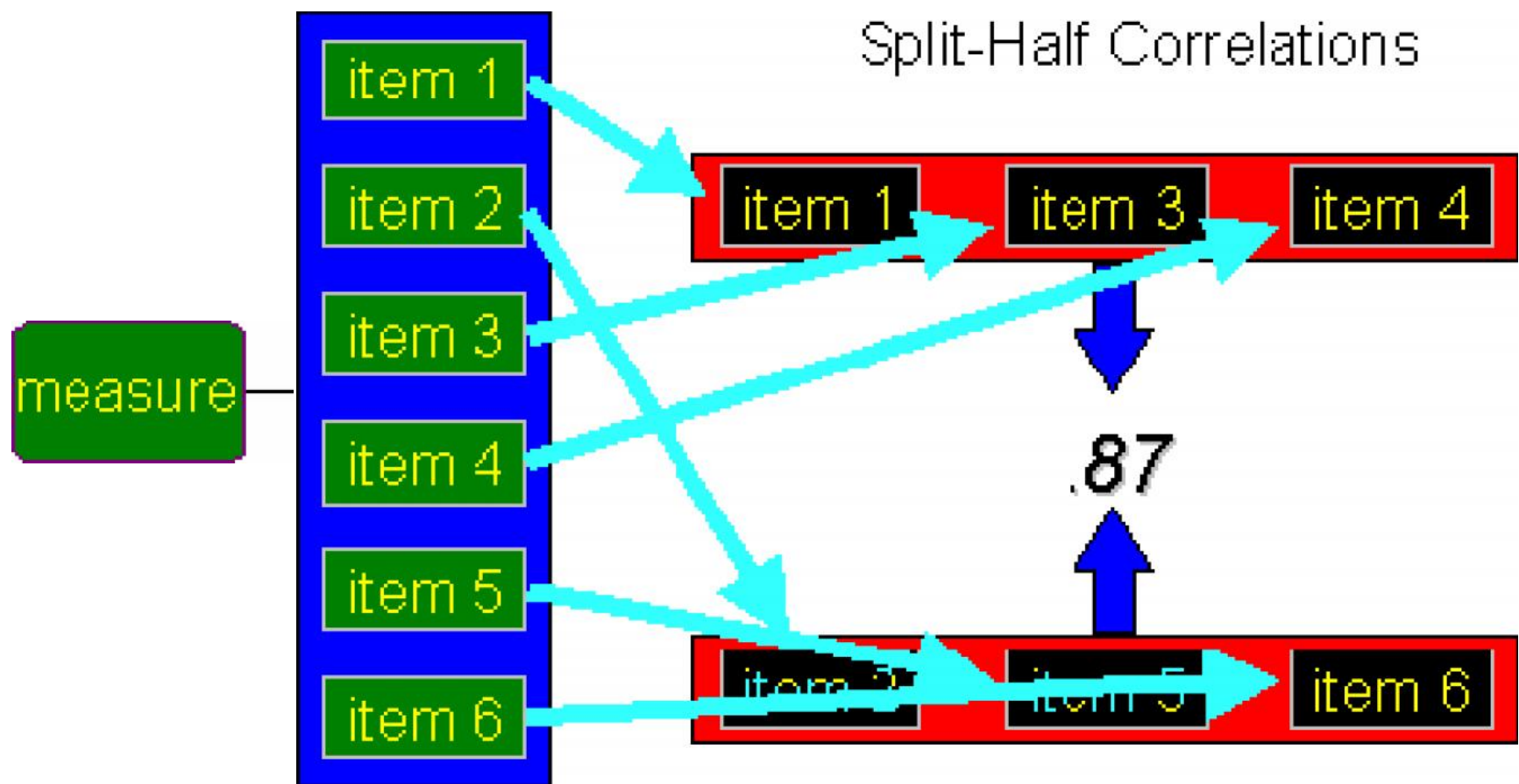
Internal Consistency Reliability

Average item-total correlation



Internal Consistency Reliability

Split-half reliability



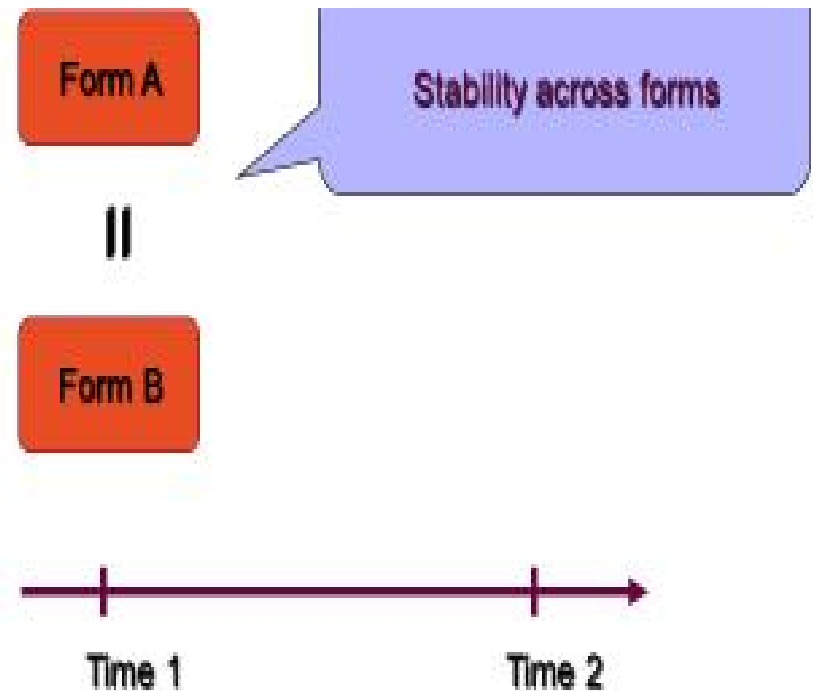
Parallel-Forms Reliability

Advantages

- Eliminates the problem of memory effect.
- Reactivity effects (i.e., experience of taking the test) are also partially controlled.

Disadvantages

- Are the two forms of the test actually measuring the same thing.
- More Expensive
- Requires additional work to develop two measurement tools.

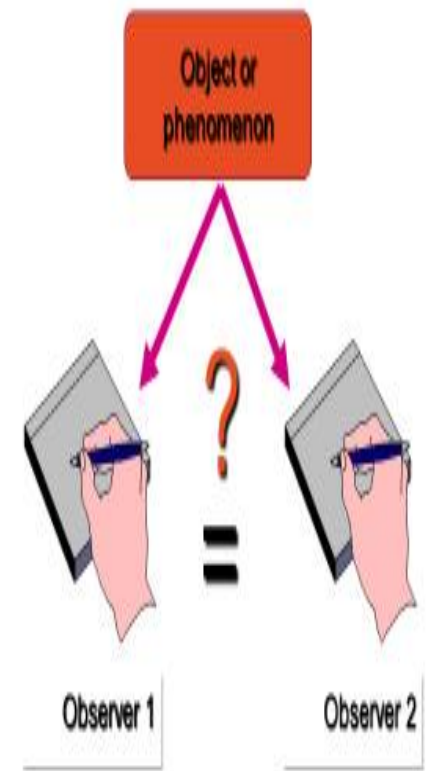


coefficient of equivalence

Inter-Rater = Inter-Observer Reliability

Inter-rater reliability refers to the consistency of measurement recorded by two or more raters or clinicians on the same cohort of patients

It establishes the equivalency (Stability and consistency across raters/examiners) of measurements and the magnitude of error associated (e.g. Large measurement error) indicated poor reliability and may suggest deviations from the standardized test protocols



Intra-Rater reliability

Intra-rater reliability refers to the consistency of measurement recorded by one rater or clinician across two or more occasions.


If the test is 100% reliable, each patient's score will be identical on both occasions of testing, assuming no clinical change in their status.

In reality, errors occur even when measurements are undertaken by one clinician. Determining the magnitude of these errors is invaluable in interpreting the results of measurement, particularly whether a patient has improved, deteriorated or remained unchanged.




Inter-Rater = Inter-Observer Reliability

There are a number of statistics that have been used to measure interrater and intra-rater reliability.

- ❖ A percent of agreement
 - ❖ Cohen's kappa (for two raters),
 - ❖ Adaptation of Cohen's kappa (3 or more raters)
 - ❖ Pearson intra-class correlation coefficient
 - ❖ Spearman intra-class correlation coefficient
- 

Factors Affecting Reliability

- 1) **Number** of items (the more questions, the higher the reliability)
 - 2) Item **difficulty** (moderately difficult items lead to higher reliability, e.g., p-value of .40 to .60)
 - 3) **Homogeneity/similarity** of item content (e.g., item x total score correlation; the more homogeneity, the higher the reliability)
 - 4) Scale format/number of response **options** (the more options, the higher the reliability)
- 

Points to Ponder

A measure CANNOT be valid but reliable

- Necessary but not sufficient
- Reliability is a prerequisite for validity
- Any reliability is only an estimate of consistency
-