Reliability

Mohammed TA, Omar Ph.D. PT

Rehabilitation Health Science Department
College of Applied Medical Sciences-KSU

Objectives

- Defines reliability and distinguish among the various types.

- Explores ways of establishing reliability and how it can be reported using descriptive and statistical meth

Reliability

Reliability

- What is reliability and its significant ?
- Types of reliability
  - Test-reteset reliability
  - Internal consistency
  - Parallel form reliability
  - Split half reliability
  - Intrarater reliability
  - Interrater-reliability

Reliability analysis

- 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

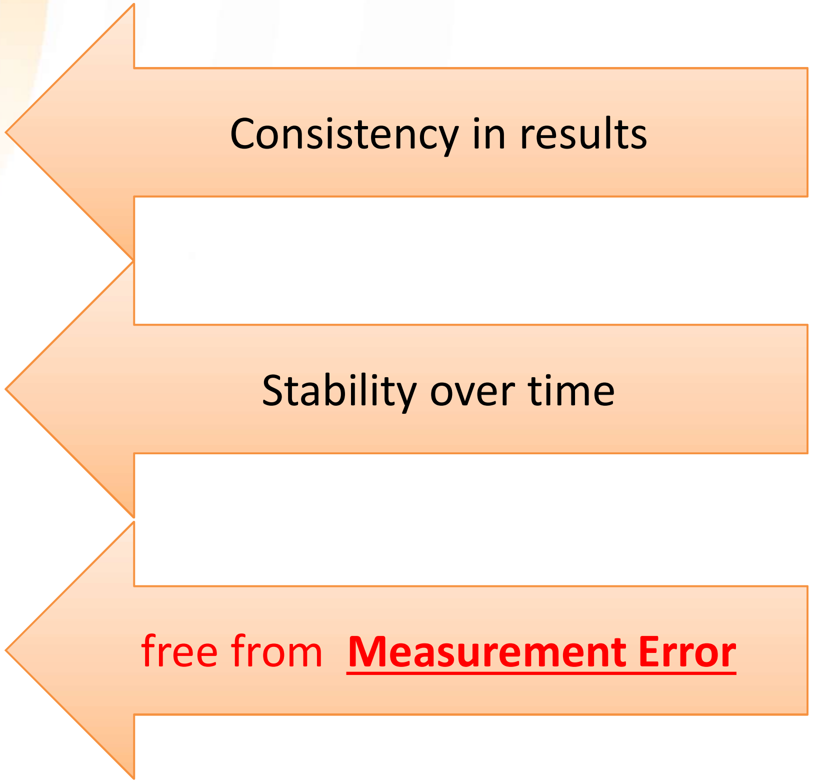
Reliability and its significant

Reliability

"…the degree to which a test or measure produces the same scores when applied in the same circumstances…"

(Nelson 1997)

What is?



Consistency in results

Stability over time

free from **Measurement Error**

Reliability and its significant

- Reliability is **“not an all-or-none”** phenomenon

Higher error   Lower error

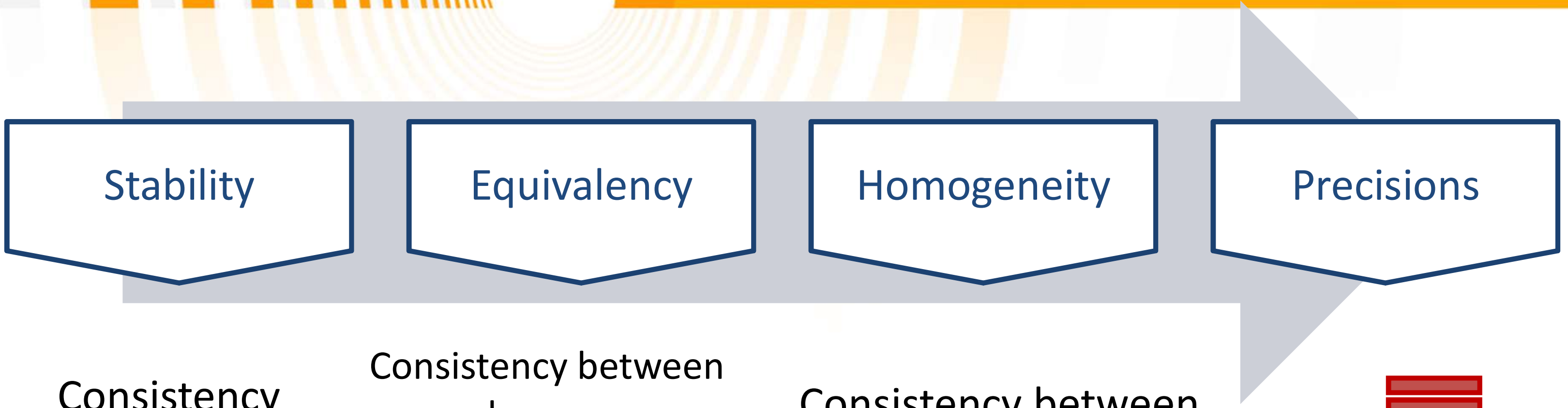
0   1

Reliability continuum

SS   OB

- Reliability is a pre-requisite of validity
- No sufficient for decisions making
- measure is never universally reliable.

Reliability and its significant



Stability

Equivalency

Homogeneity

Precisions

Consistency across time

Test-retest Reliability

Consistency between observers

Interrater/intrarater Reliability

Consistency between instrument items

Parallel form reliability

Consistency between items measures the same concept

Internal consistency reliability

=

Stability

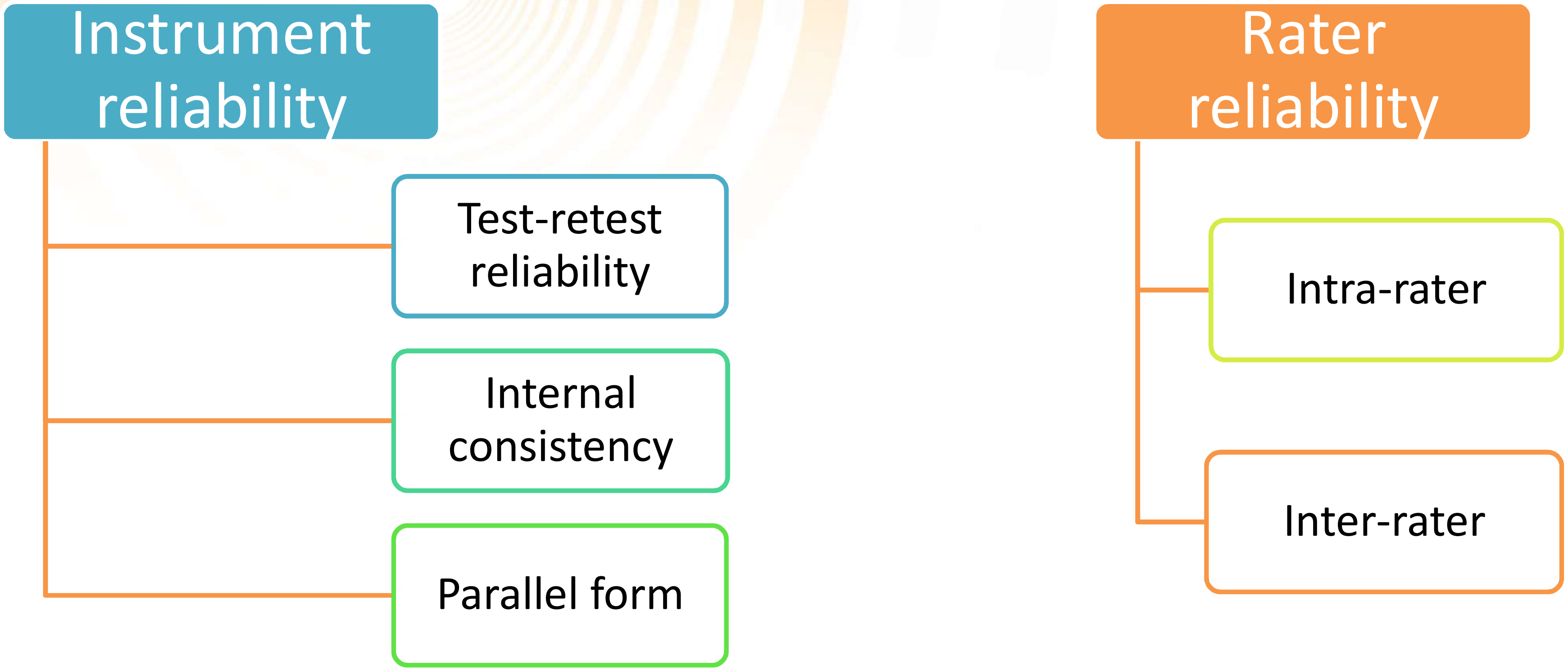
+

Equivalency

+

Homogeneity

Types of Reliability



Instrument reliability

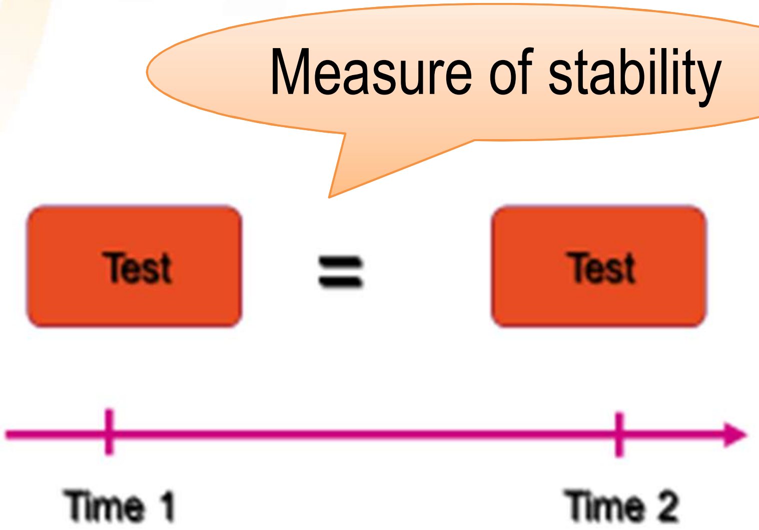
- Test-retest reliability
- Internal consistency
- Parallel form

Rater reliability

- Intra-rater
- Inter-rater

Test-Retest Reliability

- Same raters/observers
- Same groups/individual
- Used same Instrument
- At two different times.
- For PROMs and/or Performance OMs



Measure of stability

Test = Test

Time 1

Time 2

Monitor changes following treatment

Test-Retest Reliability

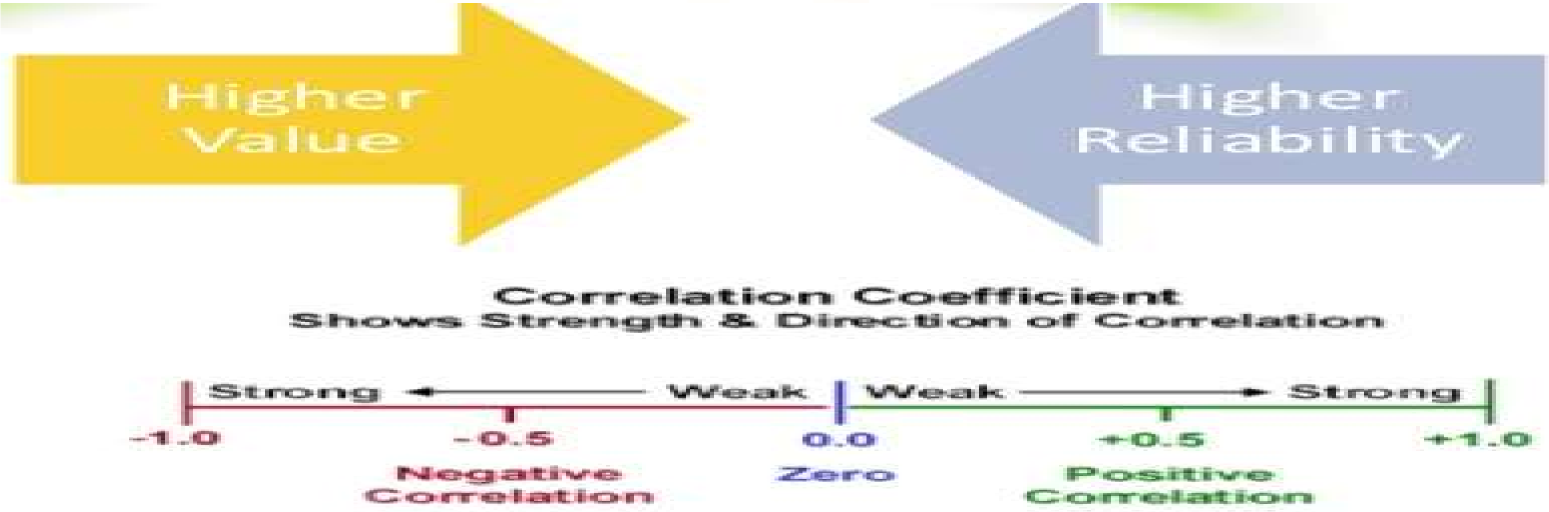
Issues should considers for test-retest reliability:

- Subject attrition between testing.
- Time laps to measures reliability (2days -6weeks)
    - Longer the time gap, the lower the test-retest reliability (construct my be change)
    - Shorter the time gap, the higher the test-retest reliability (memorization/recall)
    - Traits and actual change in health of over time
- Motivation/ fatigue
- Learning /practice effect (e.g. performance test)
- A single examiner can duplicate the results
- Interclass correlation coefficient (ICC) is the most frequently used to estimate test–retest reliability (group comparisons, ICC ≥ 0.7; individuals comparisons, (ICC ≥0.9)

Test-Retest Reliability

Issues should considers for test-retest reliability:

- ❖ Interclass correlation coefficient (ICC) is the most frequently used to estimate test–retest reliability (group comparisons, ICC ≥ 0.7; individuals comparisons, (ICC ≥0.9)



Test-Retest Reliability

Joint		Measurement (n=30)		ICC (95% CI)
		1st	2nd	
Shoulder	Flexion	147.2±16.0	152.9±14.2	0.906 (0.79–0.95)*
	Extension	50.3±12.5	51.9±13.9	0.808 (0.57–0.91)*
Hip	Flexion	107.4±12.0	107.8±11.7	0.946 (0.87–0.97)*
	Extension	23.1±7.2	24.2±8.2	0.955 (0.89–0.98)*

Test-retest reliabilities of range of motion measurements using goniometer

J Phys Ther Sci. 2016 Mar; 28(3): 722–724.

Internal Consistency Reliability

- Internal consistency describes the extent to which all the items in a test measure the same concept or construct. (correlation of test with itself).

- It is most commonly associated with PROs (paper & pencil test)

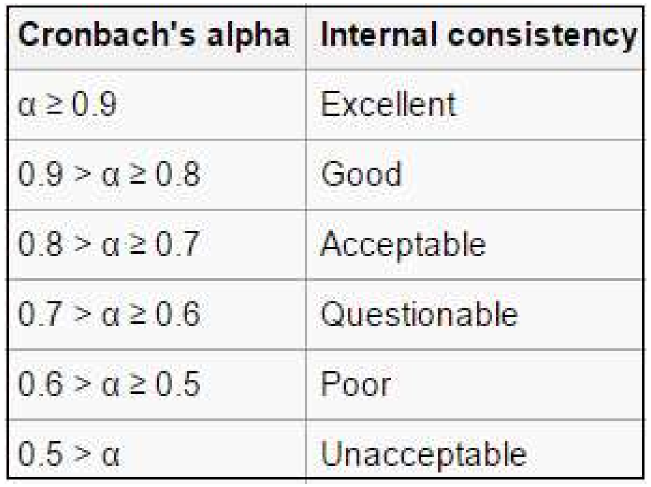
- Internal consistency is concerned with the interrelatedness of a sample of test items, not homogeneity of scale

- Internal consistency should be determined before a test can be used for research or examination purposes to ensure validity

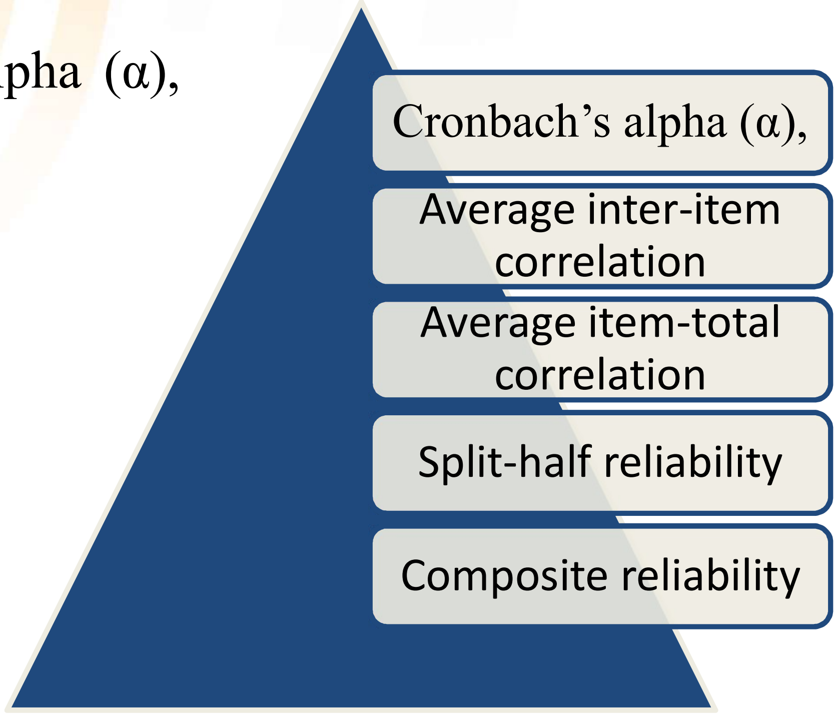
Internal Consistency Reliability

Internal consistency

Frequently evaluated with Cronbach's alpha (α), generally acceptable at values of 0.7-0.9.7



Cronbach's alpha	Internal consistency
α ≥ 0.9	Excellent
0.9 > α ≥ 0.8	Good
0.8 > α ≥ 0.7	Acceptable
0.7 > α ≥ 0.6	Questionable
0.6 > α ≥ 0.5	Poor
0.5 > α	Unacceptable



Cronbach's alpha (α),

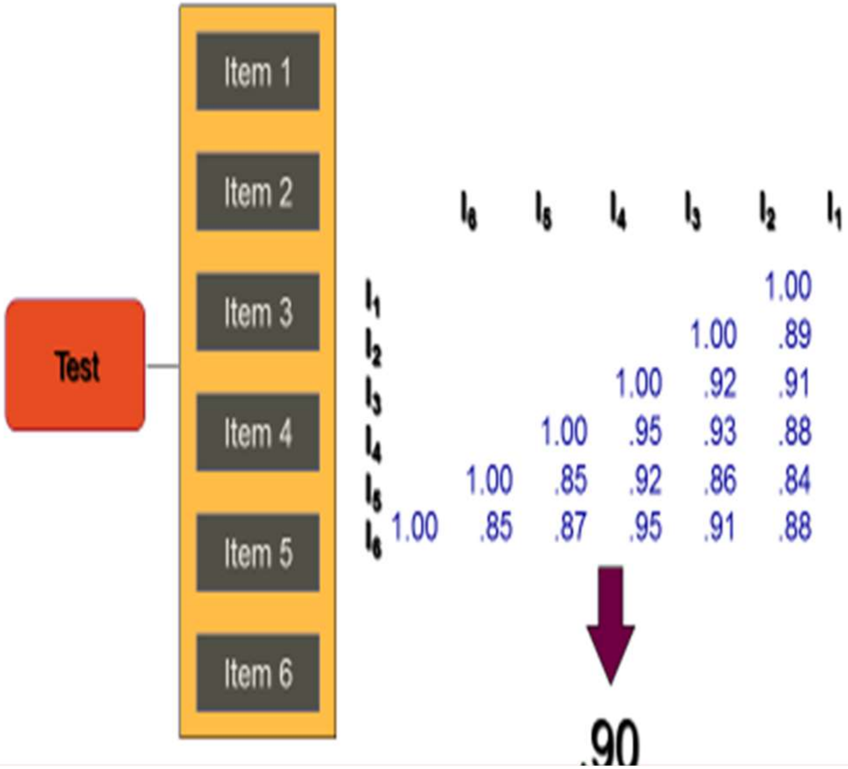
Average inter-item correlation

Average item-total correlation

Split-half reliability

Composite reliability

Average inter-item correlation



	I_6	I_5	I_4	I_3	I_2	I_1
I_1						1.00
I_2					1.00	.89
I_3				1.00	.92	.91
I_4			1.00	.95	.93	.88
I_5		1.00	.85	.92	.86	.84
I_6	1.00	.85	.87	.95	.91	.88

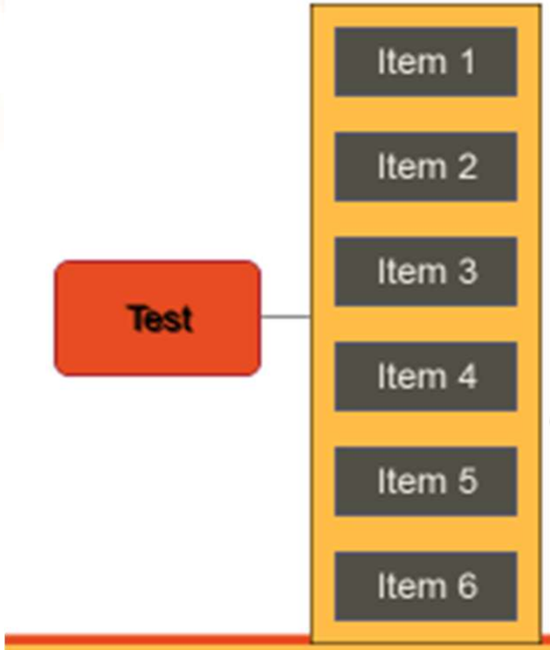
Inter-item correlations are an essential element in conducting an item analysis of a set of test questions.

Inter-item correlations examine the extent to which scores on one item are related to scores on all other items in a scale.

It provides an assessment of item redundancy: the extent to which items on a scale are assessing the same content

Average inter-item correlation should be between .20 and .40,

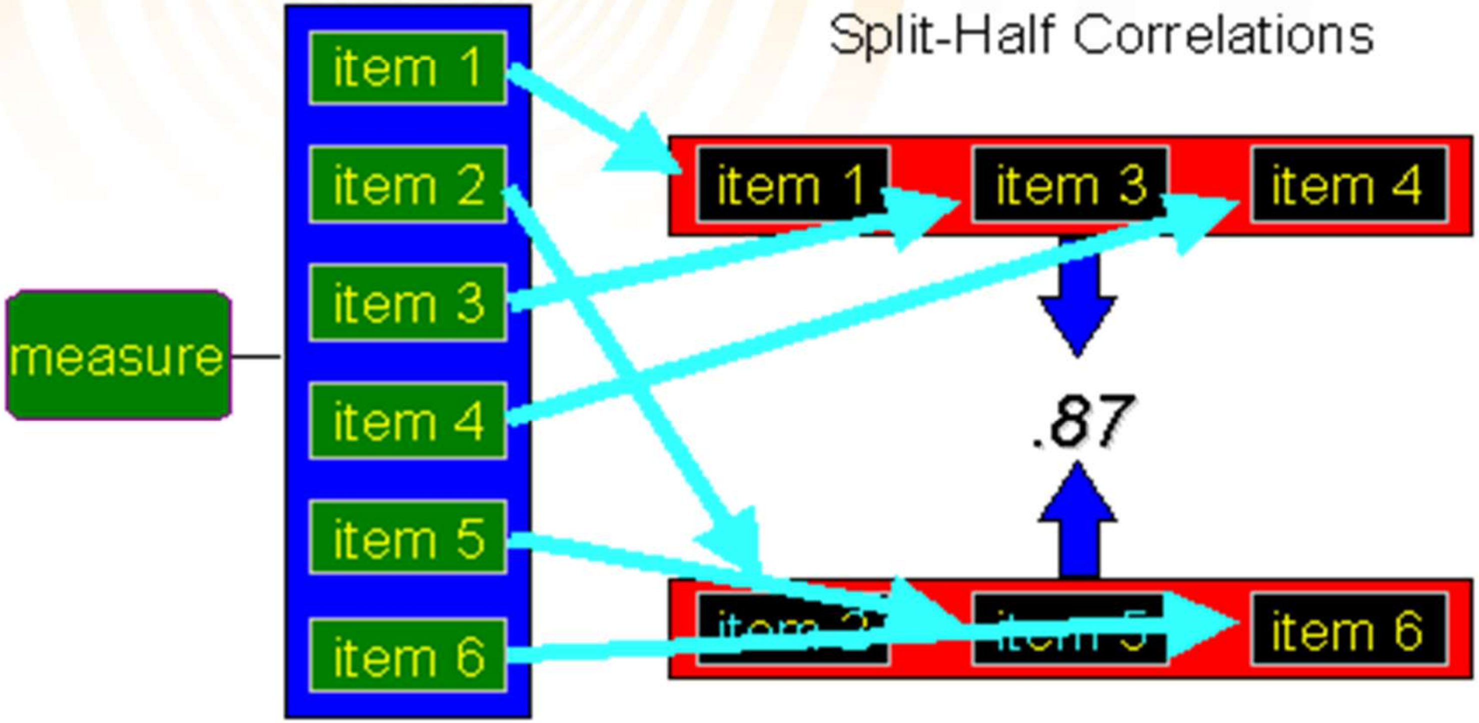
Average item-total correlation



	I_6	I_5	I_4	I_3	I_2	I_1	
I_1	1.00						
I_2	.89	1.00					
I_3	.91	.92	1.00				
I_4	.88	.93	.95	1.00			
I_5	.84	.86	.92	.85	1.00		
I_6	.88	.91	.95	.87	.85	1.00	
Total	.84	.88	.86	.87	.83	.82	1.00

.85

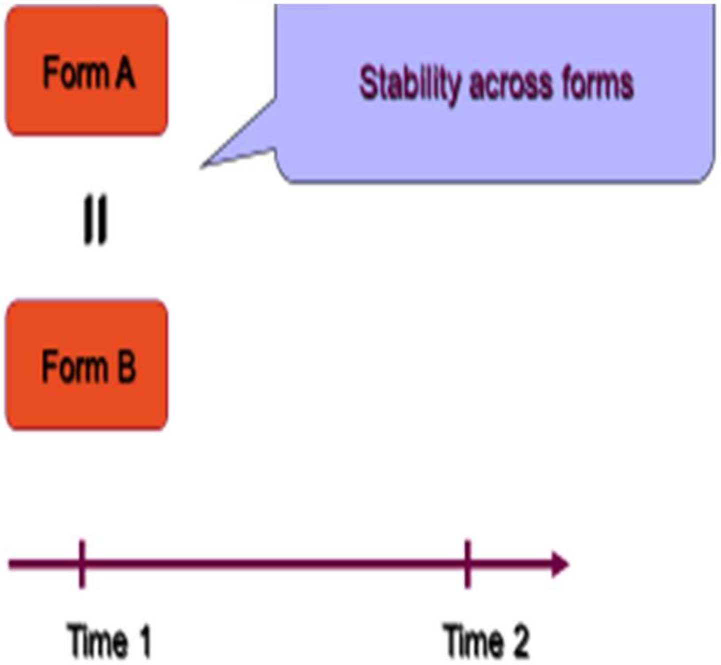
Split-half reliability



Parallel Test Reliability

- Used when development of multi-item parallel tests (alternative-form tests) is desirable.

- Parallel forms reliability indicates the consistency of responding to different item samples (the two test forms) and, when the forms are administered at different times, the consistency of responding over time.



Parallel Test Reliability

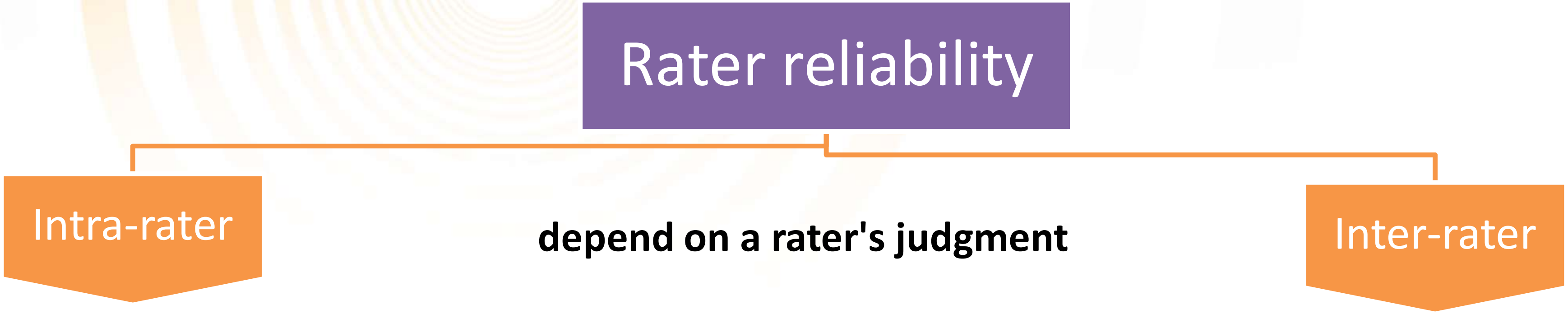
Advantage

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

Disadvantage

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

Interrater and Intrarater Reliability



Rater reliability

Intra-rater

depend on a rater's judgment

Inter-rater

- Assesses the consistency of the same rater measuring on two or more occasions, blinded to the scores he or she assigned on any previous measurements.

- Assessment involves having two or more observers independently applying the same instrument with the same people and comparing scores for consistency.

Inter-Rater 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

- High variation among individuals being tested
- Standardisation of testing Procedures
  - Clear instructions
  - Optimal testing situation
- Fatigue
- Habituation and learning effects

Factors Affecting Reliability

1) **Lengthen of test (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)

Exercises -1-

- Place the letter of the type of reliability listed in the left-hand column next to the term that best matches it in the right-hand column:

Types of Reliability	Related Terms
A. Test–Retest	____ Used when multi-item tests are needed that measure same the construct.
B. Parallel Test	____ Assesses responses from the same scorer at different times.
C. Interrater	____ Stability, Reproducibility.
D. Intrarater	____ Assesses responses from different scorers.

Designing a Reliability Study

Study Design + Types of reliability + *Timing of Measurements* + Target Population / sampling + Other Design Issues

?????

Test-Retest
Parallel Test
Interrater
Intrarater

Internal consistency

Simultaneity
Interrater reliability

interval for PRO
1-2wks

Physical performance
Short interval

Sample recruitment
Sample characteristics

A sample size of 50 is adequate in most reliability studies

3-5 participants /item

- Blinding.
- Training.
- Attrition.
- Random ordering of items or subscales.

Exercises -2-

Checking the attached files and answer the following

- Describe the scale/instrument /questionnaire used , timing of measurement, target population and sampling types of included reliability and how they are assessed and interpreter

Exercises -2-

- Validity and Reliability of the Chronic Respiratory Disease Questionnaire in Elderly Individuals with Mild to Moderate Non-Cystic Fibrosis Bronchiectasis Respiration 2015;90:89–96

- Reliability and validity of 4-metre gait speed in COPD, European Respiratory Journal 2013 42: 333-340;

Exercises -2-

- Reliability of Ashworth and Modified Ashworth Scales in Children with Spastic Cerebral Palsy BMC Musculoskeletal Disorders 2008, 9:44

- Reliability and validity of the Chinese version of the pediatric quality of life inventoryTM (PedsQLTM) 3.0 neuromuscular module in children with Duchenne muscular dystrophy Health Qual Life Outcomes. 2013; 11: 47.

Exercises -2-

- Urdu version of the neck disability index: a reliability and validity stud Farooq et al. BMC Musculoskeletal Disorders (2017) 18:149 DOI

- Neck Pain and Disability Scale and Neck Disability Index: validity of Dutch language versions Eur Spine J (2012) 21:93–100

- Cross-cultural Adaptation, Reliability, and Validity of the Arabic Version of Neck Disability Index in Patients With Neck Pain SPINE Volume 38, Number 10, pp E609–E615

Exercises -2-

- Cross-cultural adaptation and validation of the Saudi Arabic version of the Knee Injury and Osteoarthritis Outcome Score (KOOS). Rheumatology International (2018) 38:1547–1555