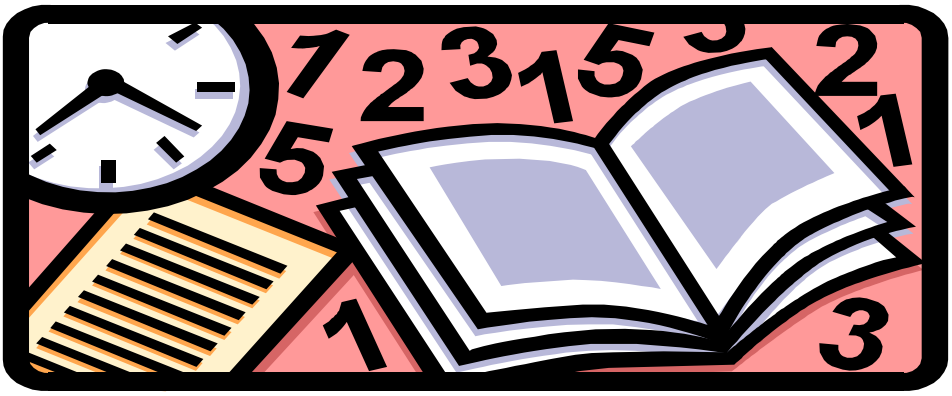


Reliability & Validity

Dr. Sudip Chaudhuri

M. Sc., M. Tech., Ph.D., M. Ed.

Assistant Professor, G.C.B.T. College, Habra, India,
Honorary Researcher, Saha Institute of Nuclear Physics,
Life Member, Indian Society for Radiation and Photochemical Sciences (ISRAPS)
chaudhurisudip@yahoo.co.in

Measurement Error

A participant's score on a particular measure consists of 2 components:

Observed score = True score + Measurement Error

True Score = score that the participant would have obtained if measurement was perfect—i.e., we were able to measure without error

Measurement Error = the component of the observed score that is the result of factors that distort the score from its true value

$x_{total} = x_{true} + x_{total}$ **and** $\sigma^2_{total} = \sigma^2_{true} + \sigma^2_{total}$

Factors that Influence Measurement Error

- **Transient states** of the participants:
  (transient mood, health, fatigue-level, etc.)
- **Stable attributes** of the participants:
  (individual differences in intelligence, personality, motivation, etc.)
- **Situational factors** of the research setting:
  (room temperature, lighting, crowding, etc.)

Characteristics of Measures and Manipulations

- Precision and clarity of operational definitions
- Training of observers
- Number of independent observations on which a score is based (more is better?)
- Measures that induce fatigue or fear

Actual Mistakes

- Equipment malfunction
- Errors in recording behaviors by observers
- Confusing response formats for self-reports
- Data entry errors

Measurement error undermines the reliability (repeatability) of the measures we use

Reliability

- The reliability of a measure is an inverse function of measurement error:
- The more error, the less reliable the measure
- Reliable measures provide consistent measurement from occasion to occasion

Estimating Reliability

Total Variance in a set of scores	=	Variance due to true scores	+	Variance due to error
Reliability	=	True-score Variance	/	Total Variance

$$r_{tt} = \frac{\sigma^2_{true}}{\sigma^2_{total}}$$

Reliability can range from 0 to 1.0

When a reliability coefficient equals 0, the scores reflect nothing but measurement error

Rule of Thumb: measures with reliability coefficients of 70% or greater have acceptable reliability

Different Methods for Assessing Reliability

- **Test-Retest Reliability**
- **Internal Consistency Reliability**

Test-Retest Reliability

- **Test-retest reliability refers to the consistency of participant's responses over time (usually a few weeks, why?)**
- **Assumes the characteristic being measured is stable over time—not expected to change between test and retest**

$$r_t = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n-1)S_x S_y} \quad \textbf{and} \quad r_{tt} = \frac{2r_t}{1 + r_t}$$

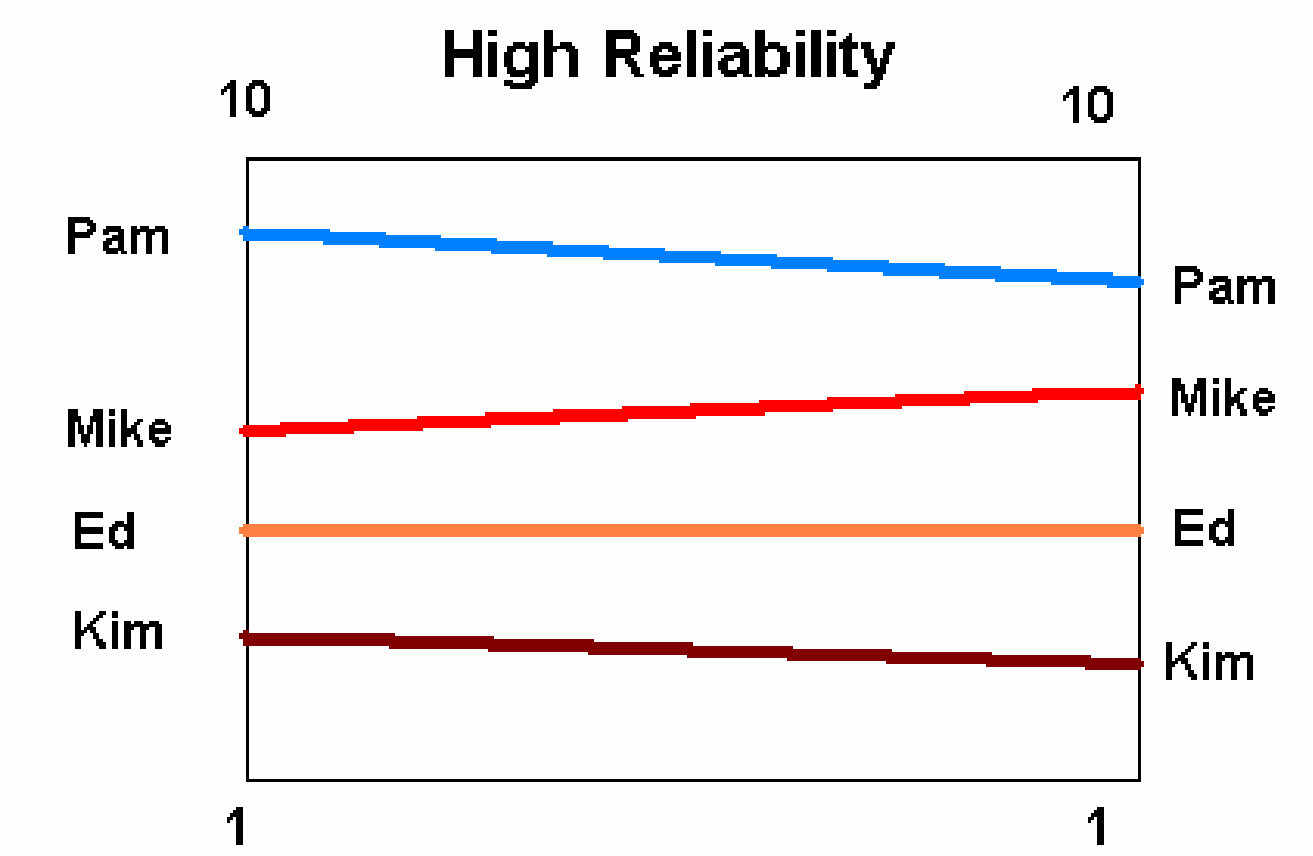
Internal Consistency Reliability

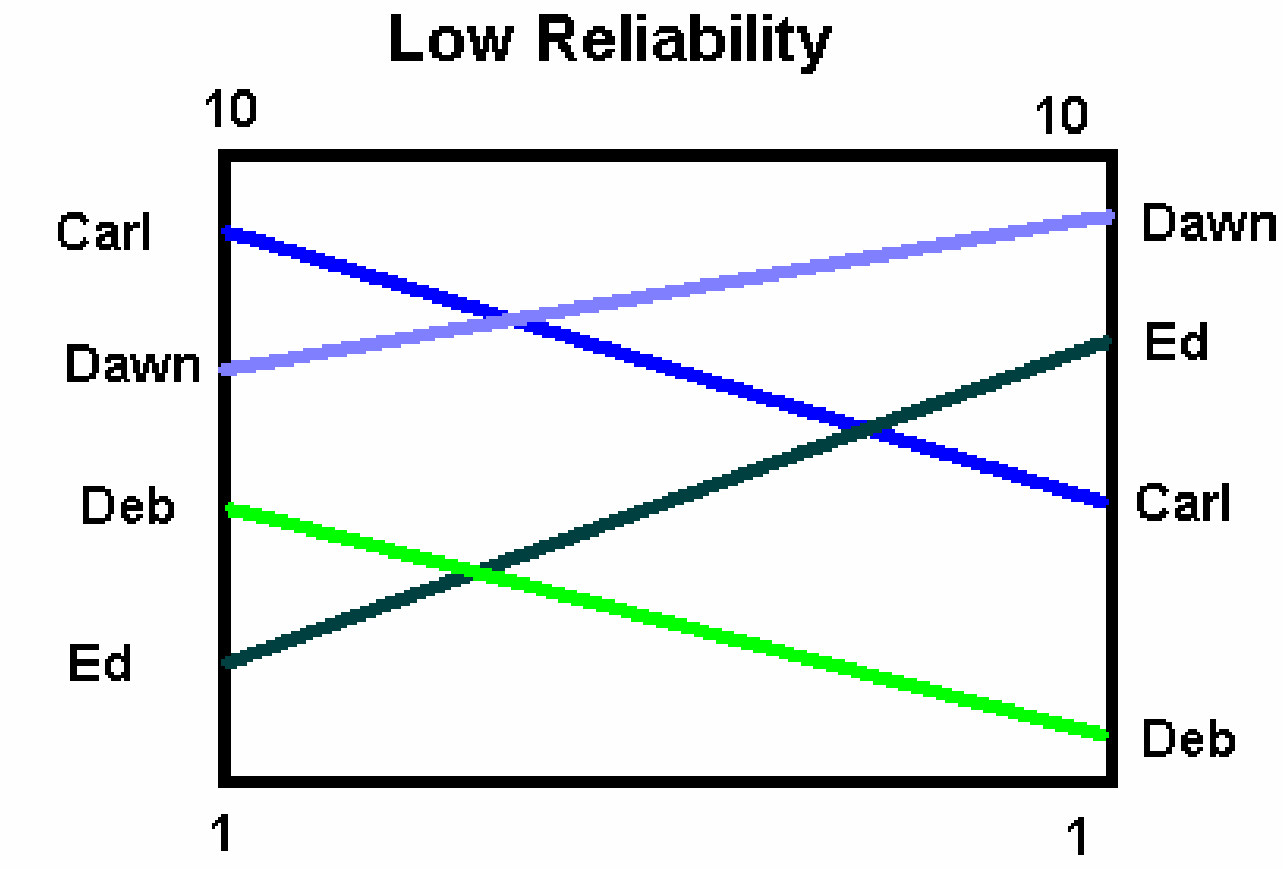
- Relevant for measures that consist of more than 1 item (e.g., total scores on scales, or when several behavioral observations are used to obtain a single score)
- Internal consistency refers to inter-item reliability, and assesses the degree of consistency among the items in a scale, or the different observations used to derive a score
- Want to be sure that all the items (or observations) are measuring the same construct

Estimates of Internal Consistency

- **Item-total** score consistency
- **Split-half reliability**: randomly divide items into 2 subsets and examine the consistency in total scores across the 2 subsets (any drawbacks?)

$$r_{tt} = \frac{n}{(n-1)}\left\{ 1 - \frac{M(n-M)}{n\sigma^2{}_t} \right\}$$


High Reliability
10
10
Pam
Mike
Ed
Kim
Pam
Mike
Ed
Kim
1
1


Low Reliability
10
10
Carl
Dawn
Dawn
Ed
Deb
Carl
Ed
Deb
1
1

Factors affecting Reliability

Extrinsic Factors:

1. Group variability (↓↓)
2. Guessing by the examinees
3. Envoronmental conditions
4. Momentary fluctuations in the examinee

Factors affecting Reliability

Intrinsic Factors:

1. Length of the test (↑↑)
2. Homogeneity of items (↑↑)
3. Facility value (FV) of test items (~0.5)
4. Discrimination index (DI) of test items (↓↓)
5. Range of total scores (↓↓)
6. Scorer reliability

How to improve Reliability?

- Quality of items; concise statements, homogenous words (some sort of uniformity)
- Adequate sampling of content domain; comprehensiveness of items
- Longer assessment – less distorted by chance factors
- Developing a scoring plan (esp. for subjective items – rubrics)
- Ensure **VALIDITY**

Estimating the Validity of a Measure

- A good measure must not only be reliable, but also valid
- A valid measure measures what it is intended to measure
- Validity is not a property of a measure, but an indication of the extent to which an assessment measures a particular construct in a particular context—thus a measure may be valid for one purpose but not another
- A measure cannot be valid unless it is reliable, but a reliable measure may not be valid

Estimating Validity

- Like reliability, validity is not absolute
- Validity is the degree to which variability (individual differences) in participant's scores on a particular measure, reflect individual differences in the characteristic or construct we want to measure
- Three types of measurement validity:

  Face Validity

  Construct Validity

  Criterion Validity

Face Validity

- Face validity refers to the extent to which a measure 'appears' to measure what it is supposed to measure
- Not statistical—involves the judgment of the researcher (and the participants)
- A measure has face validity—'if people think it does'
- Just because a measure has face validity does not ensure that it is a valid measure (and measures lacking face validity can be valid)

Construct Validity

- Most scientific investigations involve hypothetical constructs—entities that cannot be directly observed but are inferred from empirical evidence (e.g., intelligence)
- Construct validity is assessed by studying the relationships between the measure of a construct and scores on measures of other constructs
- We assess construct validity by seeing whether a particular measure relates as it should to other measures

Convergent and Discriminant Validity

- To have construct validity, a measure should both:
- Correlate with other measures that it should be related to (**convergent validity**)
- And, not correlate with measures that it should not correlate with (**discriminant validity**)

Criterion-Related Validity

- Refers to the extent to which a measure distinguishes participants on the basis of a particular behavioral criterion
- The Scholastic Aptitude Test (SAT) is valid to the extent that it distinguishes between students that do well in college versus those that do not
- A valid measure of marital conflict should correlate with behavioral observations.
- A valid measure of depressive symptoms should distinguish between subjects in treatment for depression and those who are not in treatment

Two Types of Criterion-Related Validity

- **Concurrent validity**

  measure and criterion are assessed at the same time

- **Predictive validity**

  elapsed time between the administration of the measure to be validated and the criterion is a relatively long period (e.g., months or years)

Predictive validity refers to a measure's ability to distinguish participants on a relevant behavioral criterion at some point in the future

Examples of Concurrent and Predictive Validity

- High school seniors who score high on the the SAT are better prepared for college than low scorers (**concurrent validity**)
- Probably of greater interest to college admissions administrators, SAT scores predict academic performance four years later (**predictive validity**)

Factors that can lower Validity

- **Unclear directions**
- **Difficult reading vocabulary and sentence structure**
- **Ambiguity in statements**
- **Inadequate time limits**
- **Inappropriate level of difficulty**
- **Poorly constructed test items**
- **Test items inappropriate for the outcomes being measured**
- **Tests that are too short**
- **Improper arrangement of items (complex to easy?)**
- **Identifiable patterns of answers**
- **Teaching**
- **Administration and scoring**
- **Students**
- **Nature of criterion**