Reliability: Conceptual Basis

I. Concept of reliability

II. Reliability vs. validity

III. Theoretical foundations

IV. Estimating reliability

I. Concept of Reliability

The concept of reliability is of the consistency or precision of a measure

- Weight example

Reliability varies along a continuum, measures are reliable to a greater or lesser extent

Not an all or nothing quality

I. Concept of Reliability

The opposite of consistency and precision is variability due to *random measurement error*

Reliability is lack of random measurement error

Random error is unexplained variation that is *not systematic*

If variability is random, there will be some overestimates and some underestimates

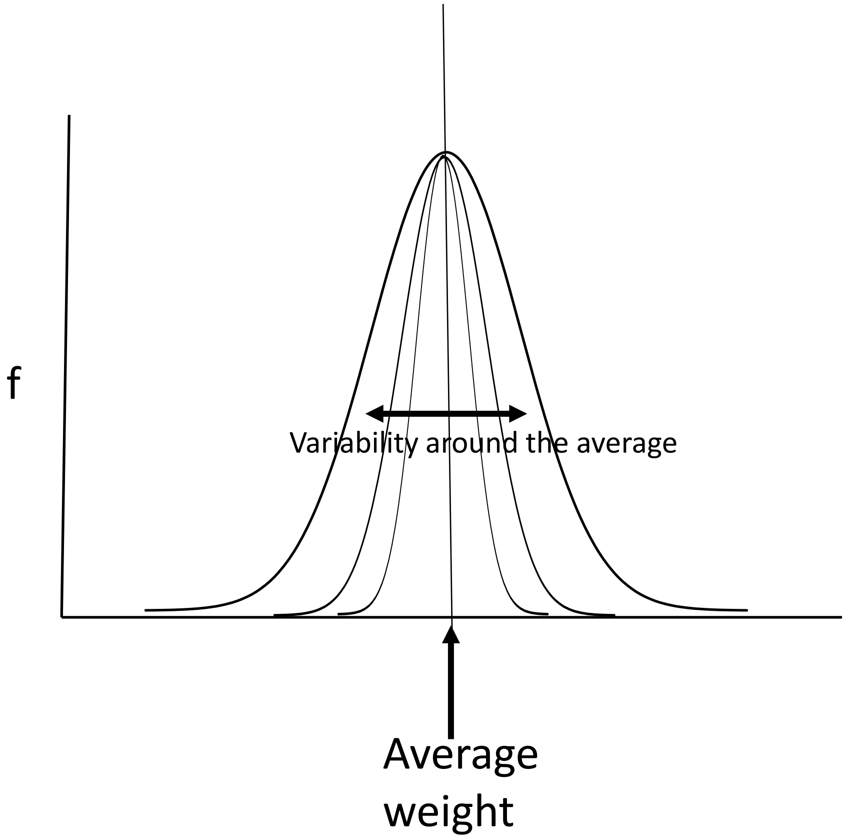
On average estimate is accurate

I. Concept of Reliability

Weight

Measurement	Weight (lbs.)
1	147
2	143
3	145
4	144
5	146
Average	*145*

I. Concept of Reliability



I. Concept of Reliability

For psychological measures, error may result from circumstances that differ in each administration

- Examples: mood, environmental noise, inconsistent testing conditions, guessing, misreading the question

If circumstances are consistent (e.g., noisy room), the effects on scores are systematic and not random

II. Reliability vs. Validity

Validity pertains to the meaning of the measure—what is the hypothetical concept or construct that the measure really captures?

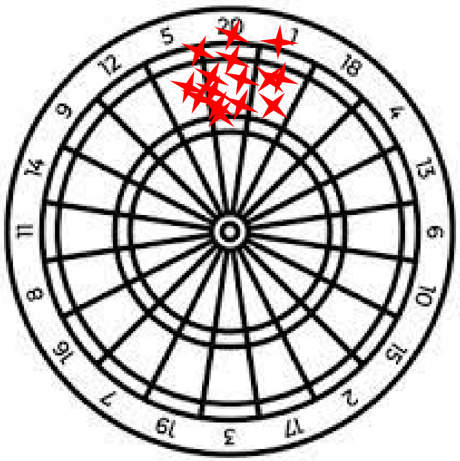
e.g., actual body weight or body weight, clothes, and heavy shoes

II. Reliability vs. Validity

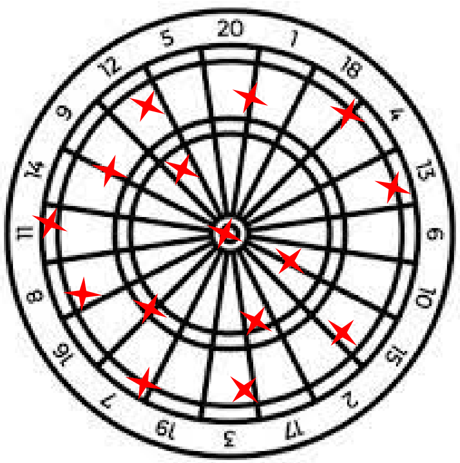
A measure is not valid to the extent that systematic variation captured is not what the researcher expects to measure

II. Reliability vs. Validity

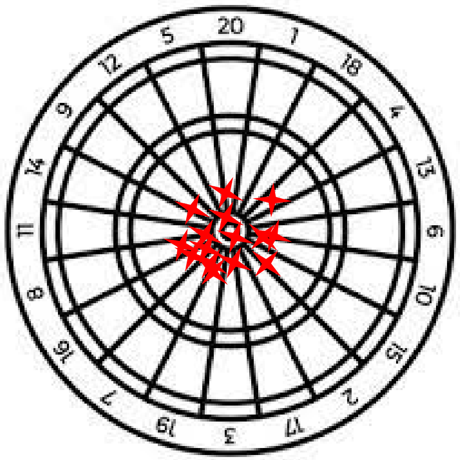
Target analogy



Reliable but not valid



Valid but not reliable



Valid and reliable

III. Theoretical Foundations

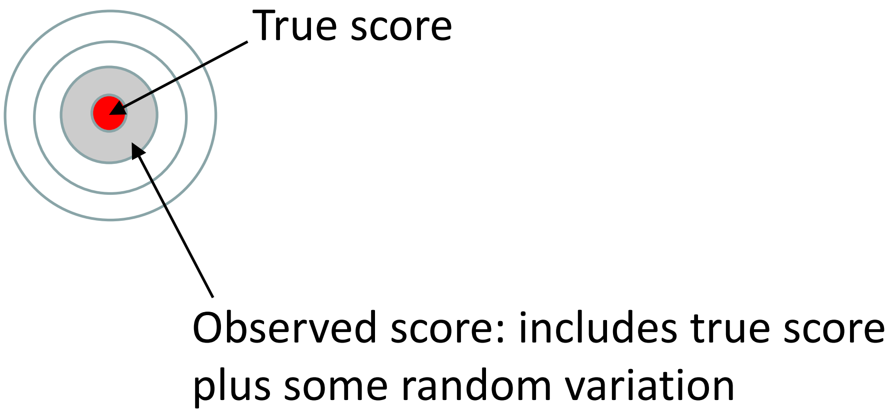
The *true score* is the correct value of the psychological attribute (construct) that we intend to measure

III. Theoretical Foundations

The *observed score* contains the true score plus other variation

Text describes “signal” plus “noise”

III. Theoretical Foundations



(Assuming no systematic variation other than true score variation)

III. Theoretical Foundations

The observed score will always have a variance as large or larger than the true score

III. Theoretical Foundations

Classical Test Theory (CTT)

Observed Score	=	*True Score*	+	*Error*
X_o	$=$	X_t	$+$	X_e

Note: many texts use $X = T + E$

III. Theoretical Foundations

Reliability is the proportion of the observed score variance, s_o^2, that is due to the true score, s_t^2

The smaller the error variance, s_e^2, the greater proportion that is due to true score variance and the higher the reliability

If proportion is 1.0, then no error variance -> perfect reliability

If proportion is 0.0, then all error variance –> no reliability and all noise

III. Theoretical Foundations

Weight

Professor	Weight (lbs.)		
	X_o	X_t	X_e
1	170	180	-10
2	195	170	25
3	145	160	-15
4	135	150	-15
5	165	140	25
6	120	130	-10
Average	155	155	0
Variance	608.33	291.67	316.67

$$s_t^2 + s_e^2 = s_o^2$$
$$291.67 + 316.67 = 608.33$$

III. Theoretical Foundations

$$\text{Reliability} = \frac{\text{True}}{\text{True} + \text{Error}}$$

$$R_{xx} = \frac{s_t^2}{s_t^2 + s_e^2}$$

$$= \frac{s_t^2}{s_o^2}$$

Note: your text uses R_{xx} as the symbol for reliability but most texts use ρ_{xx} (rho) or r_{xx}

III. Theoretical Foundations

Other ways to think about reliability

Squared correlation between observed and true score, $R_{xx} = r_{ot}^2$

One minus the squared correlation between the observed score and error, $R_{xx} = 1 - r_{oe}^2$ (one minus the proportion error)

Small standard error of measurement – average size of the error scores

IV. Estimating Reliability

Three ways to think of the standard error of measurement:

Standard deviation of measurement errors, $se_m = \sqrt{s_e^2} = s_e$

Conceptually the average error (deviation of the observed score from the true score) for repeated measurements

The degree to which the observed score has greater variability than the true score due to unreliability, $se_m = s_0\sqrt{1 - R_{xx}}$

IV. Estimating Reliability

Test-retest reliability

Repeat the test two or more times to see how similar the measurements are

Calculate the correlation between the measurement occasions

Problem is that in the interval between the measurement occasions the attribute may have changed

Small time interval needed in between measurements without contamination from recall

IV. Estimating Reliability

Parallel tests

Two tests are parallel if their true scores are the same and they have the same standard deviation

Theoretical notion, because it is not possible to know with absolute certainty that two tests are exactly parallel

IV. Estimating Reliability

Alternative forms reliability

- If we could create two parallel or alternative forms of a measure, we could estimate reliability of the measure without repeated measurements
  - e.g., standardized tests, like the SAT and GRE, use alternative test forms

IV. Estimating Reliability

Split-half reliability

- Can develop a larger test and correlate two halves
- Problem is how best to split up the test
  - e.g., what if the first half and second half differ?

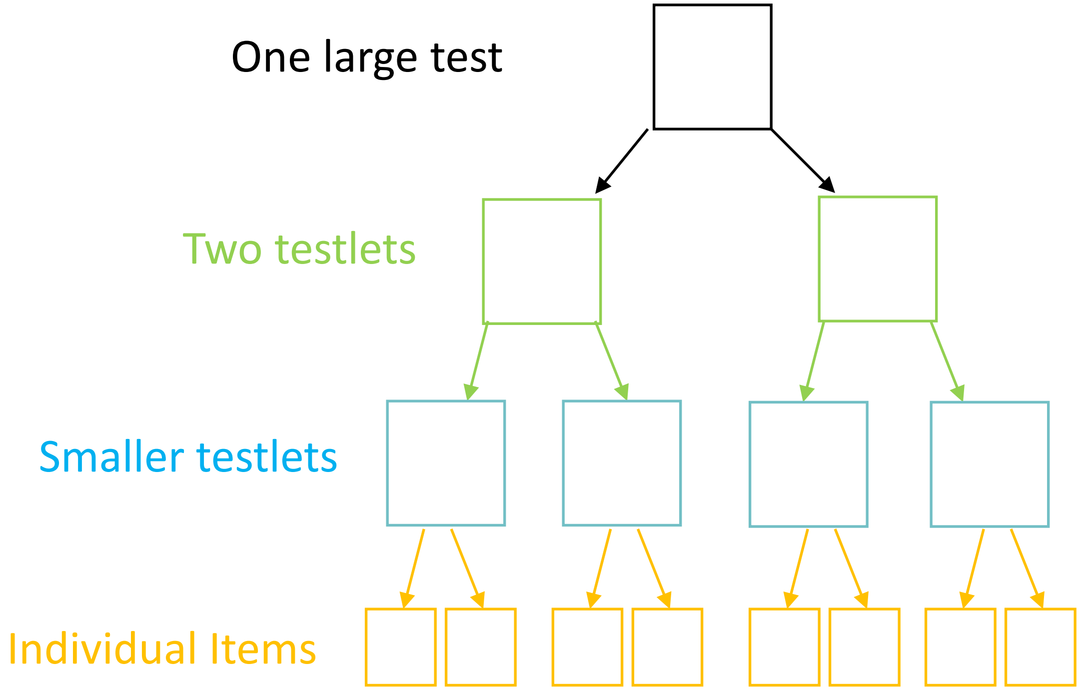
IV. Estimating Reliability

Domain sampling theory (model)

What if we considered a set of items from a test to be from a larger pool (domain, population) of items from the same test

We could think of every item as a small parallel test, a *testlet* or *subtest*

IV. Estimating Reliability



IV. Estimating Reliability

Domain sampling theory (model)

If we view each item as good representations of the true score and each as a random selected item from a domain or population of possible items, then we can relax the assumption that each test is strictly parallel

Instead we only need to think of them as on average equally representing the domain

IV. Estimating Reliability

Internal reliability

The domain sampling idea allows us to use the correlations among items to gauge the reliability of a measure

This is the basis of *internal reliability*, such as the type of reliability assessed by Cronbach's alpha

IV. Estimating Reliability

Inter-rater reliability

For observational measures, we often have two or more raters assess the same behavior

Calculating the correlation between the separate ratings assesses reliability in a similar fashion to test-retest reliability