1. represents the extent to which the results really measure what they are supposed to measure.

a. How is it assessed?

* By checking how well the results correspond to established theories and other measures of the same concept.

2. In contrast, extent to which the results can be reproduced when the research is repeated under the same conditions.

a. How is it assessed?

* By checking the consistency of results across time, across different observers, and across parts of the test itself.

3. How are validity and reliability related?

a. A measure is not always ; the results might be reproducible, but they’re not necessarily correct.

b. A measurement is generally : if a test produces accurate results, they should be reproducible.

4. Types of validity (ways to measure validity)

a. 🡪 refers to how well an experiment is done, especially whether it avoids confounding (more than one possible independent variable [cause] acting at the same time). The less chance for confounding in a study, the higher its internal validity is.

b. 🡪 validity of applying the conclusions of a scientific study outside the context of that study. In other words, it is the extent to which the results of a study can be generalized to and across other situations, people, stimuli, and times.

c. 🡪 also called logical validity, is a simple form of

validity where you apply a superficial and subjective assessment of whether or not your study or test measures what it is supposed to measure. Put differently, face validity is the degree to which a procedure, especially a psychological test or assessment, appears effective in terms of its stated aims.

d. 🡪 Evidence that a variable corresponds with (i.e., concurrent validity) or predicts (i.e., predictive validity) another variable. Criterion validity (or criterion-related validity) measures how well one measure predicts an outcome for another measure. A test has this type of validity if it is useful for predicting performance or behavior in another situation (past, present, or future).

e. 🡪 when the predictor and criterion data are collected at the same time. It can also refer to when a test replaces another test (i.e. because it’s cheaper). For example, a written driver’s test replaces an in-person test with an instructor.

f. 🡪 if the test accurately predicts what it is supposed to predict. For example, the SAT exhibits predictive validity for performance in college. It can also refer to when scores from the predictor measure are taken first and then the criterion data is collected later.

5. Ways to measure reliability

* 1. The easiest way to access reliability is to use
     1. This method randomly splits the data set into two
     2. A score for each participant is calculated based on the half of each scale
     3. If a scale is reliable a person’s score on one half should be very similar to their score on the other half

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Respondent | SPLIT #1 | | SPLIT #2 | |
| JOBSAT1 | JOBSAT2 | JOBSAT3 | JOBSAT4 |
| 1 | 7 | 6 | 3 | 6 |
| 2 | 6 | 6 | 2 | 4 |
| 3 | 6 | 5 | 4 | 2 |
| 4 | 7 | 7 | 3 | 4 |

Is this a reliable measure? Why?

* + 1. A scale is reliable if the split-half correlation is and .
  1. Another way to examine reliability is to test the same thing twice.
     1. This type of reliably check is called *test-retest reliability*.
     2. Will this reliability assessment work for all types of instruments?

* 1. Perhaps the best way to assess the reliability of an instrument is to calculate its

*Cronbach’s alpha*.

* + 1. This is loosely equivalent to splitting the data in two in every possible way and calculating the *average of each intercorrelation*.
    2. Cronbach’s alpha () ranges from 0 to 1; *.70*  is an \*acceptable\*  score.
    3. It highly influenced by the *number of items* in the scale

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SCALE #1 | | | SCALE #2 | | |
| # of items |  | Average r  between items | # of items |  | Average r  between items |
| 3 | .8 | .57 | 10 | .8 | .28 |

Threats to validity

There are numerous threats to the validity of cause and effect conclusions (see Cook, Campbell, & Peracchio [1990] in Supplemental Materials content folder)

* 1. Reverse causation 🡪

* 1. Reciprocal causation 🡪

* 1. Spurious relation 🡪

* 1. Moderated relation 🡪

* 1. Selection of participants 🡪

* 1. Mortality 🡪

* 1. Compensatory rivalry 🡪

*\*\*Interesting fact: Saretsky (1972) named this the “John Henry” effect in honor of the steel driver who, upon learning that his output was being compared with that of a steam drill, worked so hard that he outperformed the drill and died of overexertion.*

* 1. Resentful demoralization 🡪

* 1. Hypothesis guessing 🡪