Measurement: Reliability and Validity

Y520
Strategies for Educational Inquiry

Robert S Michael
Introduction: Reliability & Validity

- All measurements, especially measurements of behaviors, opinions, and constructs, are subject to fluctuations (error) that can affect the measurement’s reliability and validity.

- **Reliability** refers to consistency or stability of measurement.
  - Can our measure or other form of observation be confirmed by further measurements or observations?
  - If you measure the same thing would you get the same score?

- **Validity** refers to the suitability or meaningfulness of the measurement.
  - Does this instrument describe accurately the construct I am attempting to measure?

- In statistical terms:
  - Validity is analogous to unbiasedness (valid = unbiased)
  - Reliability is analogous to variance (low reliability = high variance)
Reliability

- The observed score on an instrument can be divided into two parts:
  - Observed score = true score + error
  - An instrument is said to be reliable if it accurately reflects the true score, and thus minimizes the error component.
  - The reliability coefficient is the proportion of true variability to the total observed (or obtained) variability.
  - E.g., If you see a reliability coefficient of .85, this means that 85% of the variability in observed scores is presumed to represent true individual differences and 15% of the variability is due to random error.

- Reliability is a correlation computed between two events:
  - Repeated use of the instrument (stability)
  - Similarity of items (split half / internal consistency)
  - Equivalence of 2 instruments (equivalence)
Stability: Test-Retest

- Test - retest reliability: Same group of respondents complete the instrument at two different points in time. How stable are the responses?
  - The correlation coefficient between the 2 sets of scores describes the degree of reliability.
  - For cognitive domain tests, correlations of .90 or higher are good. For personality tests, .50 or higher. For attitude scales, .70 or higher.

- Problems with test-retest reliability procedures:
  - Differences in performance on the second test may be due to the first test - i.e., responses actually change as a result of the first test.
  - Many constructs of interest actually change over time independent of the stability of the measure.
  - The interval between the 2 administrations may be too long and the construct you are attempting to measure may have changed.
  - The interval may be too short. Reliability is inflated because of memory.
Alternate (aka Equivalent) Forms

- Involves using differently worded questions to measure the same construct.
  - Questions or items are reworded to produce two items that are similar but not identical.
  - Items must focus on the same exact aspect of behavior with the same vocabulary level and same level of difficulty.
  - Items must differ only in their wording.
- Reliability is the correlation between the responses to the pairs of questions.
- Alternate forms reliability is said to avoid the practice effects that can inflate test-retest reliability (i.e., respondent can recall how they answered on the identical item on the first test administration).
Alternate Forms: Example

- From the Watson-Glaser Critical Thinking test:

  - **Test A**: Terry, don’t worry about it. You’ll graduate someday. You’re a college student, right? And all college students graduate sooner or later.

  - **Test B**: Charlie, don’t worry about it. You’ll get a promotion someday. You’re working for a good company, right? And everyone who works for a good company gets a promotion sooner or later.
Internal Consistency: Homogeneity

■ Is a measure of how well related, but different, items all measure the same thing.
■ Is applied to groups of items thought to measure different aspects of the same concept.
■ A single item taps only one aspect of a concept. If several different items are used to gain information about a particular behavior or construct, the data set is richer and more reliable.
Internal Consistency Example

Example:

- The Rand 36-item Health Survey measures 8 dimensions of health. One of these dimensions is “physical function.”
- Instead of asking just one question, “How limited are you in your day-to-day activities?” Rand found that asking 10 questions produced more reliable results, and conveyed a better understanding of “physical function.”
Internal Consistency: Rand Example

The following questions are about activities you might do during a typical day. Does your health now limit you in these activities. If so, how much? (Response options are: limited a lot, limited a little, not limited at all).

- Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports.
- Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf.
- Lifting or carrying groceries.
- Climbing several flights of stairs.
- Climbing one flight of stairs.
- Bending, kneeling, or stooping.
- Walking more than a mile.
- Walking several blocks.
- Walking one block.
- Bathing or dressing yourself.
Internal Consistency: Cronbach’s Alpha

- Cronbach’s alpha:
  - Indicates degree of internal consistency.
  - Is a function of the number of items in the scale and the degree of their intercorrelations.
  - Ranges form 0 to 1 (never see 1).
  - Measures the proportion of variability that is shared among items (covariance).
  - When items all tend to measure the same thing they are highly related and alpha is high.
  - When items tend to measure different things they have very little correlation with each other and alpha is low.
  - The major source of measurement error is due to the sampling of content.
Cronbach’s Alpha

- Conceptual Formula:

\[ \alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}} \]

- where \( k \) = number of items and \( r \) = the average correlation among all items.

- Given a 10 item scale with the average correlation between items is .50 (e.g., calculate the correlation between items 1 and 2, between item 1 and 3, between 1 and 4, etc. for all possible pairs). (10 taken 2 at a time = 45 pairs)

- \( \alpha = 10 \times (.5) / (1+9(.5)) = 0.91 \)
Validity

- Validity indicates how well an instrument measures the construct it purports to measure.
- Types of validity:
  - Construct validity
  - Content validity
  - Criterion validity
Content Validity

- Is a judgment of how appropriate the items seem to a panel of reviewers who have knowledge of the subject matter.
  - Does the instrument include everything it should and nothing it should not?
  - An instrument has content validity when its items are randomly chosen from the universe of all possible items. Example: Constructing a vocabulary test using a sample of all vocabulary words studied in a semester.
  - Content validity is a subjective judgment; there is no correlation coefficient.
  - A table of specification helps.
Criterion Validity

- Criterion validity indicates how well, or poorly, and instrument compares to either another instrument or another predictor.
  - It has two parts: concurrent validity & predictive validity
  - **Concurrent** validity requires that the scale be judged against some other method that is acknowledged as a “gold standard” (e.g., another well-accepted scale). If the correlation between the new scale and the old scale is strong, the new scale is said to have concurrent validity.
  - **Predictive** validity is the ability of a scale to forecast future events, behaviors, attitudes or outcomes – Do SAT scores predict first semester GPA?
  - Validity coefficients are correlation coefficients between the new scale and the gold standard or some future outcome.
Construct Validity

- Indicates how well the scale measures the construct it was designed to measure.
- Ways of establishing construct validity:
  - Known group differences
  - Correlation with measure of similar constructs
  - Correlation with unrelated & dissimilar constructs
  - Internal consistency
  - Response to experimental manipulation
  - Opinion of judges.
Inferences in Measurement

- The following slide contains a list of nine items.
- For each item:
  - Describe how we measure,
  - Identify the level of measurement,
  - Describe the inference we make, and
  - Describe the construct.
Inferences in Measurement

- Length of table
- Weight of person
- Speed of car
- Temperature
- Humidity
- Wind chill index
- Discomfort Index (Smog Index / Pollen Index)
- Intelligence
- Anxiety
Inventing Constructs

- How would you describe the relationship(s) among the following:
  - Operational definitions
  - Variables
  - Measurement
  - Scales of measurement
  - Inferences from measurements
  - Constructs

- Discuss each for the items on the following slide
Inventing Constructs

- Temperature
- Wind Chill
- Discomfort Index
- Intelligence
- Extroversion
Example:

Intelligence and its Measurement

- Intelligence is one construct that continues to generate discussion, and divergent opinions.
- The *Journal of Educational Psychology* held a symposium in 1921 entitled “Intelligence and its Measurement.”
- Definitions of the constructed were offered by researchers. As you read, how would you operationalize each definition? What would items on a scale look like?
Definitions:

Intelligence and its Measurement (a)

- The ability to give responses that are true or factual. — E. L. Thorndike
- The ability to carry on abstract thinking. — L. M. Terman
- The ability to learn to adjust oneself to the environment. — S. S. Colvin
- The ability to adapt oneself to relatively new situations in life. — R. Pinter
Definitions:

Intelligence and its Measurement  

- The capacity for knowledge and knowledge possessed. — V. A. C. Henmon
- A biological mechanism by which the effect of a complexity of stimuli are brought together and given a somewhat unified effect in behavior. — J. Peteson
Definitions:

Intelligence and its Measurement (c)

- The ability to inhibit an instinctive adjustment, to redefine the inhibited instinctive adjustment in the light of imaginally experienced trial and error, and to realize the modified instinctive adjustment into overt behavior to the advantage of the individual as a social animal. — L. L. Thurstone
Definitions:
Intelligence and its Measurement (d)

- The ability to acquire abilities. — H. Woodrow
- The ability to learn or profit by experience. — W. F. Dearborn

“Intelligence is what intelligence tests test.”
— E. G. Boring (1923)