Measurement

Validity & Reliability
What is measurement validity?

- Validity refers to the accuracy of a measure.
- A measurement is valid when it measures what it is suppose to measure and performs the functions that it purports to perform.
- Does an indicator accurately measure the variable that it is intended to measure?
Given the imperfect nature of measurement, validity is a matter of degree.
Approaches to validity include:

- Content
- Face
- Predictive
- Concurrent
- Construct
Content

- Determined by expert judgements of the appropriateness of the contents of a measure.
Content validity refers to degree that one has representatively sampled from that domain of meaning.

This type of validity receives the most attention in construction of achievement & proficiency measures with psychology & education.
For Example: I want to assess the content validity of a test in research methods.

- Consider test time & types of items.
- There are 20 topics I wish to cover and I decide to use 35 multiple choice in 50 minute period.
- I might have one question on each topic & distribute remaining questions to topics I consider most important.
I also need to consider the skill level to measure.

For example, knowledge of facts & definitions, application of definitions, drawing inferences, and making critical appraisals, and so forth.
Face Validity

• Determined by judgements made by the researcher and based on surface appearance.
There are two basic types:
- Predictive
- Concurrent
In both predictive and concurrent validity, we validate by comparing scores with a criterion.
What is a criterion?

- This is the standard by which your measure is being judged or evaluated.
Predictive Validity

- This is determined by correlating test scores with criterion scores obtained after examinees have had a chance to perform what is predicted by the test.
- A measure is validated with reference to future standing on criterion variable.
What is the criterion

- A measure of what the test is designed to predict.
When is the criterion measured?

- After examinees have had a chance to exhibit the predicted behavior
Example: Employment Test

- The purpose of an employment test is to predict success on the job.
- The most appropriate test of its validity is predictive validity - to what extent does the test predict what it is supposed to predict?
How do we do it?

- Give test to applicants for a position.
- For all those hired, compare their test scores to supervisors’ rating after 6 months on the job.
- The supervisors’ ratings are the criterion.
- If employees scored on the test similarly to supervisors’ ratings, then predictive validity of test is supported.
Concurrent Validity

- Correlate test scores with criterion scores obtained at about the same time.
- The ability of a measure to indicate an individual’s present standing on the criterion variable.
What is the criterion?

• An independent measure of the same trait that the test is designed to measure.
When is the criterion measured?

- At about the same time that the test is administered.
Example: Reading Achievement Test

- A multiple choice test on reading achievement (which is designed to measure achievement at the time of testing and not designed to predict any future behavior.) might be validated by:
  - comparing the scores on the test with teachers’ ratings of students’ reading abilities.
- Teachers’ ratings are the criterion
Other examples:

- Validate a measure of political conservatism by correlating it with reported voting behavior.
- Travis Hirschi validated his index of self-reported delinquency by comparing responses to office police records on the respondents.
Construct Validity

- Hypothesize a relationship between the test scores and scores on another variable.
- Then, test the hypothesis
What is a construct?

- A construct stands for a collection of related behaviors that are associated in a meaningful way.
- Depression is a construct that stands for a personality trait that is manifested by behaviors such as lethargy, loss of appetite, difficulty in concentrating on tasks, and so forth.
Those behaviors are indicators of depression.

- The construct itself does not have a physical being outside of its indicators.
- Thus, we infer its existence by observing the collection of related indicators.
- Emphasis on collection crucial because any one sign may be associated with several constructs.
Loss of appetite is a sign of depression. But loss of appetite may also be a sign of anxiety, or fear, or falling in love. Thus, loss of appetite is indicative of depression only when it is found in association with other indicators of depression.
So how do you determine the construct validity?

• Begin by hypothesizing about how the construct that the measure is designed to measure should relate to other variables.
Example:

• Hypothesize that students who are very depressed will earn lower grades & be more likely to drop out of college than students who are less depressed.
• Test the hypothesis on sample of students.
• Suppose we find no relationship between scores obtained on depression measure and success in college.
Either (1) the measure lacks validity for measuring depression; that is, it measures something else which is not related to success in college, or (2) the hypothesis is wrong. If we continue to our faith in the hypothesis, we will have to conclude that the empirical evidence argues against the validity of the measure.
Now let’s suppose that we find a relationship between scores obtained on the new depression measure and success in college.
What does this mean?

- Either, (1) the depression measure is, to some degree, valid for measuring depression, or
- (2) the depression index measures a variable other than depression that is also related to success in college.
- This “other variable” could be many things
Example of the “other variable”

- Maybe the index is heavily laden with signs of depression that are also signs of anxiety such that it is more a measure of anxiety than depression.
- Since debilitating anxiety may lead to failure in college, the scores on the scale may relate to success in college because it measures anxiety and not depression.
What is measurement reliability?
Reliability concerns:

• Is the indicator measuring ‘something’ consistently & dependably.
• Do repeated applications of the operational definition under similar conditions yield consistent results?
Approaches to reliability:

- Test-retest
- interitem (internal consistency)
- alternate-forms
- interobserver
Test-retest

- You apply a measure to a sample of people & then, at a somewhat later time, apply the same measure to the same people again.
- Result will be two sets of scores.
- Correlate the scores & should find high degree of association.
Issues to consider in establishing time frame for test-retest:

- If too short an interim, respondents may remember & simply repeat the responses they gave the first time, thereby inflating the reliability estimate.
- If too long an interim, real change in behavior may have occurred between the two tests.
Interitem (internal consistency)

- Appropriate when multiple items are used to measure a concept.
- Concern is with assessing the reliability of a set of measures of the same concept by comparing all possible combinations of these items.
- Degree to which this has been achieved is assessed by a statistic (e.g., Cronbach’s alpha) which is the average inter-item correlation for a set of items.
Alternate-forms

- Compare the responses of subjects’ to variations on a particular survey question.
- The variations on the questions should be extremely close.
- The two sets of responses should be quite similar.
- Split-halves reliability is similar to this.
Interobserver

- Using a particular measure, two or more different observers should provide similar ratings to the same phenomenon.
- If they do so, the interobserver reliability of the measure is supported.
“Some social researchers have claimed that we can assume that members of certain minority groups are more active in criminal behavior due to the fact that they are more likely to be arrested for such activities. Is this claim more easily attacked due to problems of reliability or validity?
Validity

- “Arrest rates are not necessarily indicative of rates of involvement in illegal behavior. They can be affected by police bias, location, and other demographic factors.”
- Therefore, we know that there are problems in the validity of arrest rates to function as measure of certain criminal behaviors and for certain populations.
Have researchers assessed the validity of arrest rates as measure of crime?

- National Crime Victimization Survey (NCVS) developed to provide alternate source of data on crime.
- NCVS used to assess validity of arrest rates & reports as measures of crime.
- Convergence between the two in terms of relevant factors supports official measures as more valid indicators of certain crimes (e.g., homicide) than of others (e.g., burglary).
Another example from Shutt disk

• “Joe and Mary were interested in whether male or female college students tended to have more sexual partners. They each stationed themselves at different entrances to the college cafeteria and stopped every tenth person who was entering the building and asked them how many people they had had sexual intercourse with. When they examined their results, they found that the men that Mary interviewed had had fewer partners than the men that Joe interviewed. Is this a problem with validity or reliability?”
Reliability

• “It looks like when Mary conducts the interviews, males under-report their number of partners. When {Joe} conducts the interviews, males over-report their number of partners. When two people get different results, this is a reliability problem.”