

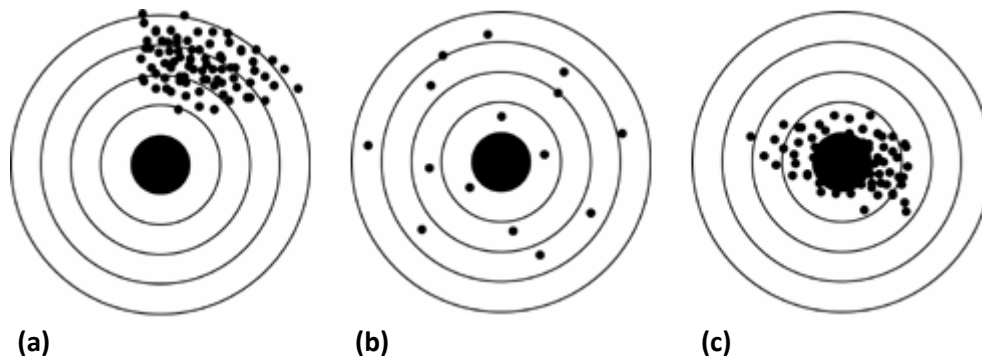
# Validity & Reliability

---

Validity refers to the quality of a measure (or study). Does it truly measure what it is intended to measure? Does an IQ test truly measure intelligence? Does a personality test truly measure one's personality? Does a pulmonary function test truly measure pulmonary function?

Reliability refers to consistency. Would the measure (or study) give us the same result if we repeated it numerous times? If a person took an IQ test several times within a short time period, would he receive the same IQ score. Since IQ shouldn't change on a daily basis, a reliable test should yield the same results when repeated.

Imagine the true measure to be the center of a target. We aim for the bull's-eye. A measure that is perfectly valid and perfectly reliable will hit the bull's-eye every time. Which of the following represent a measure that is valid but not reliable? Reliable but not valid? Both valid and reliable?



Our goal is to design and implement a measure (or study) that has high validity and high reliability. It is possible to have one with the other, and sometimes there is some give and take between the two. In order to increase validity, you may have to sacrifice reliability. Consider an English exam, for example. Using essay questions rather than multiple choice responses would increase the validity because it eliminates the possibility of random guessing. However, it would decrease reliability because of the subjectivity involved in grading essays.

## *Types of Validity*

There are several different types of validity:

1. **External Validity.** External validity refers to how well your results generalize to a larger population. A sample that is representative of the population of interest has high external validity.
2. **Internal Validity.** Internal validity is used to describe a study that involves a cause-and-effect relationship. It refers to how well the study was designed and carried out so that one can conclude it was the treatment that caused the response, not other outside factors.

3. **Face Validity.** Face validity refers to how well a study appears to be designed. It involves using your common sense and judgment to determine how well the study measures what it is intended to measure.
4. **Content Validity.** Content validity refers to how well a measure represents all facets of a concept. It involves making a checklist of criteria that should be met, then comparing the actual measure (or study) to the checklist.
5. **Criterion Validity.** Criterion validity refers to how useful a measure is as an indicator of a specific trait, either now or in the future. A measure (or study) with high criterion validity should be able to predict future outcomes and should correlate with other instruments that have been designed to measure the same thing.
6. **Construct Validity.** Construct validity refers to how well an instrument measures a theoretical concept (something that cannot be directly observed), called a construct. Examples include intelligence, creativity, and motivation.

### *Assessing Validity*

There are several questions to consider when assessing validity. Some measures are considerably more difficult than other to assess.

#### **1. External Validity**

- Were the subjects chosen randomly? Is the sample representative of the population of interest? Was the study conducted in a variety of places, using a variety of people, at different times?
- Was any pretesting done that could have influenced the responses of the subjects? Were there multiple interventions? Could earlier interventions have influenced the effects of later interventions?

#### **2. Internal Validity**

- Are groups similar in makeup at the beginning of the study? Is there a control group to account for confounding variables (variables other than the ones being studied that could influence the responses)?
- Were repeated measures taken? Did any events occur between measures that could have influenced responses?
- Were the data collected over a short period of time or a long period of time? If the data were collected over a long period of time, are there other factors that could have influenced responses?
- Were measures and collection methods consistent throughout the study? (Researches may become more skilled at observations, or may become fatigued over time.)
- Did any participants drop out of the study?
- Is the study blind? Could there have been a placebo effect (could participants have responded differently because they knew they were part of a study)?

- Could the responses be the result of regression towards the mean? If participants were originally higher-than-average scores, for example, retests will usually produce lower scores (closer to the population average).

### 3. Face Validity

- Does the study seem to be valid? Are there any obvious flaws in the measurement or design of the study?

### 4. Content Validity

- Are all facets of the concept measured?
- If an instrument of measurement is used, consider each item on the instrument. Is the skill measured by the item (1) essential, (2) useful (but not essential), or (3) unnecessary to the concept?

### 5. Criterion Validity

- Does the measure correlate with other instruments that have been designed to measure the same thing?
- Can the measure predict future behaviors?
- Can the measure distinguish between groups that should be dissimilar?

### 6. Construct Validity

- Does the measure correlate with other instruments that have been designed to measure the same thing?
- Does the measure show no correlation with other instruments that do not measure the same thing?

## *Assessing Reliability*

Reliability not only results from the measure itself (for example, clarity of questions), but also external factors (for example, background noise). Because reliability involves replication, however, it is difficult to assess reliability *before* conducting a study. One way to assess reliability before conducting a full-scale study is to pilot test your instruments so that you can get feedback from your participants regarding how the testing environment affected their performance.

### **Inter-Rater Reliability**

Inter-Rater reliability is high if two (or more) observers rate the same phenomenon in the same way. For example, a grading rubric for an essay would have high inter-rater reliability if several different teachers gave the same essay the same score. If raters are using a checklist, you can determine inter-rater reliability by calculating the percent of agreement between the raters. If the measure is quantitative, you can determine inter-rater reliability by calculating correlation between the ratings of the two observers.



### Test-Retest Reliability

Test-Retest reliability is high if there is a correlation between the same test given to the same sample at two different times. The amount of time allowed between measures is critical, since changes can occur over time.

### Parallel-Forms Reliability

Parallel-Forms reliability is high if two different versions of the same measurement given to the same sample have a high correlation.

### Internal Consistency Reliability

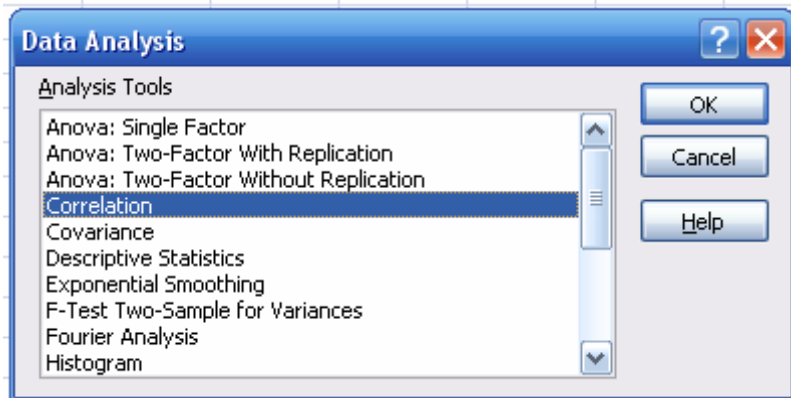
Internal Consistency reliability is high if the items on an instrument that measure the same construct are correlated. For example, on a learning-style inventory, there may be multiple questions used to determine whether a person is a visual learner. If the respondent answers each of these questions in a similar manner, the instrument is reliable. Consider the following five questions from a 25-question learning-style inventory, in which the participant must rate how much he agrees with the statement (on a scale from 1 to 5).

Item #	Question
5	I prefer to see information written on a chalkboard and supplemented by visual aids and assigned readings.
8	I can easily understand and follow directions on a map.
10	I think the best way to remember something is to picture it in your head.
20	I like to write things down or to take notes for visual review.
24	I can understand a news article better by reading about it in the newspaper than by listening to a report about it on the radio.

Although they are not exactly the same questions, they are intended to measure the same construct (the extent to which a student is a visual learner).

One way to determine internal consistency reliability is to construct an inter-item correlation matrix (assuming the data is quantitative and that the underlying relationship between any two items is linear). This matrix gives the correlation between each pair of items. For example, how does item #5 on the questionnaire correlate with item #8? If they both measure the same construct, then they should show a strong correlation. Low correlations, on the other hand, could indicate that an item is unreliable. You can construct an inter-item correlation matrix using Microsoft Excel (you will need to install the Data Analysis ToolPak).

To construct an inter-item correlation matrix using Excel, create a two-way table with each item in a different row, and each participant in a different column. From the Data Analysis menu, choose Correlation.



Highlight the rows and columns containing the data. Do not highlight the column labels. Check “Labels in first column” if you highlighted the row labels; uncheck it if you did not. Select “Rows” and hit “OK.”

	A	B	C	D	E	F
1	Item #	Subject A	Subject B	Subject C	Subject D	
2	Question 5	4	3	1	4	
3	Question 8	2	5	5	4	
4	Question 10	4	4	5	4	
5	Question 20	5	4	1	3	
6	Question 24	4	1	1	2	

The analysis shows only the lower half of the matrix since it is symmetric (the correlation between questions 5 and 20 is the same as the correlation between questions 20 and 5).

	<i>Question 5</i>	<i>Question 8</i>	<i>Question 10</i>	<i>Question 20</i>	<i>Question 24</i>
<i>Question 5</i>	1.000				
<i>Question 8</i>	-0.667	1.000			
<i>Question 10</i>	-0.943	0.471	1.000		
<i>Question 20</i>	0.828	-0.690	-0.878	1.000	
<i>Question 24</i>	0.667	-1.000	-0.471	0.690	1.000

Note that each question is perfectly correlated (1.000) with itself. Be on the lookout for correlations with an absolute value lower than 0.6; this could indicate reliability issues. Also, be on the lookout for items with a negative correlation. Does it make sense that higher responses for one question are associated with lower responses for the other? We should remove or reword questions with low inter-item reliability.

Inter-item correlation is not the only measure of internal consistency reliability. Other measures include Cronbach's alpha, Spearman's rho, and Kendall's tau.