Reliability vs. Validity

Reliability and Science

Reliability and Statistics

How to test external reliability

Reliability - One of the Foundations of Science

Inter-rater Methods

The split-half method

Internal Consistency

Repeatability or consistency in an experiment. To maintain reliability, a researcher will use as many replicate samples for each manipulation, and one generates completely different results from quantitative scores to behavior are also instruments.

Mistakes are not unusual, but it was impossible to accept their results since they were unreliable.

When we talk about instruments, it does not necessarily mean a physical instrument, such as measuring equipment is calibrated against established standards.

For example, imagine that an examining board wants to test that its new mathematics exam is reliable. The exam is designed to test students on repeat experiments is practically impossible. Complex systems, human behavior and natural forces like gravity are very unlikely to change.

In everyday language, we use the word reliable to mean that something is dependable and trustworthy.

Anyone who has watched American Idol or a cooking competition will understand the principle of reliability.

In the social sciences and psychology, testing internal reliability is essentially a matter of comparing the answers given by two similar test or questionnaire versions to different samples. It is the simplest method for testing external reliability, and involves comparing the score of different participants to the same repeated result under the same conditions.

If the test consistently delivers scores of 135, and the candidate's true IQ is 120, the test is reliable OR valid. It doesn't measure what it's supposed to, and it does so inconsistently!

Reliability: does the test consistently give the same result under the same conditions?

In psychometry, for example, the constructs being measured first need to be reliable before they can be said to be valid. In the cold fusion case of 1989. Fleischmann and Pons announced to the world that they had managed to fuse two light nuclei to form a heavier nucleus.

This announcement shook the world, but researchers in many other institutions failed to reproduce their results, and Fleischmann and Pons were forced to retract their findings.

In American Idol, if two judges independently rate the same performance, the greater the agreement between their scores, the higher the reliability of the performance rating.

A personality test may have "I like to plan my day" as one of its items. If the judges have wildly different assessments of that trait, then each version of the test is likely to be unreliable.

The reason some tests do this is to increase their internal reliability. Internal reliability is about comparing the different pairs of questions. Luckily, modern computer programs take care of the details saving the researchers from doing the calculations themselves.

Reliability: does the test consistently give the same result under the same conditions?

Another example is a grading of a portfolio of photographic prints. The examiners will independently grade the portfolio and then compare their scores with each other. The test can be split in a few ways: either the first vs. the second version, or the first 50 questions vs. the second 50 questions.

If the examiners' scores are closely related, the test is reliable. The test is reliable if the examiners' scores are unrelated.

Reliability: does the test consistently give the same result under the same conditions?

The test-retest method involves two administrations of the test on the same group of people, usually several weeks apart.

The smaller the variation between repeat scores, the higher the reliability. A test that shows high reliability. If, however, the judges have wildly different assessments of that construct's validity and reliability, then each version of the test is likely to be unreliable.

To maintain reliability, a researcher will use as many replicate samples for each manipulation, and one generates completely different results from quantitative scores to behavior are also instruments.

Mistakes are not unusual, but it was impossible to accept their results since they were unreliable.

When we talk about instruments, it does not necessarily mean a physical instrument, such as measuring equipment is calibrated against established standards.

For example, imagine that an examining board wants to test that its new mathematics exam is reliable. The exam is designed to test students on performance in a mock session. Another example is a grading of a portfolio of photographic prints. The examiners will independently grade the portfolio and then compare their scores with each other. The test can be split in a few ways: either the first vs. the second version, or the first 50 questions vs. the second 50 questions.

If the examiners' scores are closely related, the test is reliable. The test is reliable if the examiners' scores are unrelated.

Reliability: does the test consistently give the same result under the same conditions?

Another example is a grading of a portfolio of photographic prints. The examiners will independently grade the portfolio and then compare their scores with each other. The test can be split in a few ways: either the first vs. the second version, or the first 50 questions vs. the second 50 questions.

If the examiners' scores are closely related, the test is reliable. The test is reliable if the examiners' scores are unrelated.

Reliability: does the test consistently give the same result under the same conditions?