Definition of Reliability [1]

Assisted Self-Help [2]320.4K reads

Reliability is the degree of consistency of a measure. A test will be reliable when it gives the same repeated result under the same conditions.

In everyday language, we use the word reliable to mean that something is dependable and that it will give behave predictably every time. We might talk of a football player as reliable, meaning that he gives a good performance game after game.

Reliability and Science

In science, the idea is similar, but the definition is much narrower. Reliability is a property of any measure, tool, test or sometimes of a whole experiment. It's an estimation of how much random error might be in the scores around the true score.

For example, you might try to weigh a bowl of flour on a kitchen scale. A reliable scale will show the same reading over and over, no matter how many times you weigh the bowl. There may be slight error here and there – you may notice that some readings differ by just a fraction of a gram – but overall the scale is reliable. If the scale gave a reading of 1 kg and then a minute later gave a reading of 1.5 kg, the error has become so large that the instrument's reliability is seriously undermined.

When we talk about instruments, it does not necessarily mean a physical instrument, such as a mass-spectrometer or a pH-testing strip. An educational test, questionnaire [3], or assigning quantitative scores to behavior are also instruments.

Another way of looking at reliability is by considering it as a way to maximize the inherent repeatability or consistency [4] in an experiment. To maintain reliability, a researcher will use as many repeat sample groups as possible, to reduce the chance of an abnormal sample group skewing the results. This is a little like weighing the bowl several times and using the average reading.

Reliability can be determined statistically by calculating the correlation coefficient. If a test is reliable it should show a high positive correlation between repeat scores. If you use three replicate samples for each manipulation [5], and one generates completely different results from the others, there is likely something wrong with the experiment [6].

For most experiments of natural phenomena, results follow a normal distribution [7] and there is always a chance that your sample group produces results at one of the extremes. Using multiple sample groups will smooth out these extremes and generate a more accurate spread of results. But if your results continue to be wildly different, then there is likely something wrong with the design [8] itself. In this case, the entire experiment is *externally* unreliable.

External Reliability and Cold Fusion

Good experimental design will allow for plenty of replicate samples by the researchers. But other researchers should also be able to perform exactly the same experiment, with similar equipment, under similar conditions, and achieve exactly the same results. If they cannot, then the design [9] is externally unreliable.

A good example of a failure to apply the definition of reliability correctly is provided by the cold fusion case of 1989. Fleischmann and Pons announced to the world that they had managed to generate heat at normal temperatures, instead of the huge and expensive tori used in most research into nuclear fusion.

This announcement shook the world, but researchers in many other institutions failed to replicate [10] the experiment. It's unclear whether the researchers lied or genuinely made a mistake, but it was impossible to accept their results since they were unreliable.

Internal Reliability and Personality Tests

If you've ever completed a long questionnaire, you might have noticed how some questions seem to be subtle variations on one another. A personality test may have "I like to plan my activities ahead of time", "I am spontaneous" and "I like to go with the flow" as three separate items which seem quite similar.

The reason some tests do this is to increase their internal reliability. Internal reliability is about the consistency across separate items within a measure. A test is internally consistent if each item contributes equally to the overall construct being measured.

Reliability and Statistics

If you are a physicist or a chemist, repeat experiments should give exactly or almost exactly the same results, time after time. The behavior of phosphorous atoms, DNA molecules or natural forces like gravity are very unlikely to change.

Ecologists and social scientists, on the other hand, understand that achieving identical results on repeat experiments is practically impossible. Complex systems, human behavior and biological organisms are subject to far more random error and variation.

While any experimental design must attempt to eliminate confounding variables [11] and natural variations, there will always be some disparities in these disciplines.

The key to performing a good experiment [12] is to make sure that your results are as reliable as possible; if anybody repeats the experiment, statistical tests [13] will be able to compare the results and the scientist can then make a solid estimate of statistical reliability [14].

Reliability vs. Validity

Reliability and validity [15] are often confused; the terms describe two inter-related but completely different concepts. Very simply:

> *Validity: does the test actually measure what it's supposed to?*
>
> *Reliability: does the test consistently give the same result under the same conditions?*

This difference is best described with an example:

A researcher devises a new test that measures IQ more quickly than the standard IQ test:

- If the test consistently delivers scores of 135, and the candidate's true IQ is 120, the test is reliable but not valid.
- If the new test delivers scores for a candidate of 87, 65, 143 and 102, then the test is not reliable OR valid. It doesn't measure what it's supposed to, and it does so inconsistently!
- If the scores are 100, 111, 132 and 150, then the validity and reliability are also low. However, the distribution of these scores is slightly better than above, since it surrounds the true score instead of missing it entirely. Such a test is likely suffering from extreme random error.
- If the researcher's test delivers a consistent score of 118, then that's pretty close, and the test can be considered both valid and reliable. The closer to 120, the more valid, and the smaller the variation between repeat scores, the higher the reliability. A test that routinely underestimates IQ by two points can be as useful as a more valid test since the error itself is so reliable.

Reliability is an essential component of validity [16] but, on its own, is not a sufficient measure of validity. A test can be reliable but not valid, whereas a test cannot be valid yet unreliable. A test that is extremely unreliable is essentially not valid either. A bathroom scale that measures your weight one day as 5000 kg and the next day as 2 kg is not unreliable, it merely is not measuring what it is meant to.

Assessing Reliability

There are several methods to assess the reliability of instruments [17].

How to test internal reliability

In the social sciences and psychology, testing internal reliability is essentially a matter of comparing the instrument [17] with itself.

The split-half method

How could you determine whether each item on an inventory is contributing to the final score equally? One technique is the split-half method which cuts the test into two pieces and compares those pieces with each other. The test can be split in a few ways: either the first vs. the second half, or the odd-numbered items vs. the even-numbered, for example.

Split-half methods can only be done on tests measuring one construct – for example an extroversion subscale on a personality test. Psychometrics use split-half methods to identify items on a test which don't correlate strongly with the others, and then remove or improve those items.

Internal Consistency

The internal consistency test [4] compares two different *versions* of the same instrument, to ensure that there is a correlation [18] and that they ultimately measure the same thing.

For example, imagine that an examining board wants to test that its new mathematics exam is reliable, and selects a group of test students. For each section of the exam, such as calculus, geometry, algebra and trigonometry, they actually ask two questions, designed to measure the aptitude of the student in that particular area.

If there is a high internal consistency, i.e. the results for the two sets of questions are similar, then each version of the test is likely to be reliable. The test - retest method [19] involves two separate administrations of the same instrument, while internal consistency measures two different versions at the same time. Researchers may use internal consistency to develop two equivalent tests to later administer to the same group.

A statistical formula called Cronbach's Alpha [20] tests the reliability and compares various pairs of questions. Luckily, modern computer programs take care of the details saving researchers from doing the calculations themselves.

How to test external reliability

There are two common ways to establish external reliability: test-retest and inter-rater methods.

Test-Retest Method

The Test-Retest Method [19] is the simplest method for testing external reliability, and involves testing the same subjects once and then again at a later date, then measuring the correlation between those results. A test retaken after a month, for example, should yield the same results as the original, if it’s a reliable test.

One difficulty with this method lies with the time between the tests. This method assumes that nothing has changed in the meantime. If the tests are administered too close together, then participants can easily remember the material and score higher on the second round. But if administered too far apart, other variables can enter the picture: participants themselves may change enough to make their scores on the second batch not truly comparable with the first. To prevent learning or recency effects, researchers may administer a second test that is different but equivalent to the first.

Inter-rater Methods

Anyone who has watched American Idol or a cooking competition will understand the principle of inter-rating reliability. Here, what is being measured is performance, but with a panel of judges in the role of “instrument.”

An example is clinical psychology role play examinations, where students are rated on their performance in a mock session. Another example is a grading of a portfolio of photographic work or essays for a competition.

Processes that rely on expert rating of performance or skill are subject to their own kind of

error, however. Inter-rater reliability is a measure of the agreement of concordance between two or more raters in their respective appraisals, i.e. the degree of consensus among judges.

The principle is simple: if several expert raters all agree on a performance rating, that rating shows high reliability. If, however, the judges have wildly different assessments of that performance, their assessments show low reliability. Importantly, reliability is a characteristic of the ratings, and not the performance being rated.

Reliability - One of the Foundations of Science

As we have seen, understanding the definition of reliability [21] is extremely important for any scientist but, for social scientists, biologists and psychologists, it's a crucial foundation of any research design [8]. In psychometry, for example, the constructs being measured first need to be isolated before they can be measured. Thus, building inventories and tests cannot be done without constant assessment of that construct's validity and reliability. For this reason, extensive research programs always involve comprehensive pre-testing, ensuring that the instruments used are both consistent and valid.

Those in the physical sciences also perform instrumental pre-tests, ensuring that their measuring equipment is calibrated against established standards.

Links
[1] https://staging.explorable.com/en/definition-of-reliability
[2] https://staging.explorable.com/en
[3] https://explorable.com/survey-research-design
[4] https://explorable.com/internal-consistency-reliability
[5] https://explorable.com/independent-variable
[6] https://explorable.com/experimental-research
[7] https://explorable.com/normal-probability-distribution
[8] https://explorable.com/research-designs
[9] https://explorable.com/design-of-experiment
[10] https://explorable.com/reproducibility
[11] https://explorable.com/confounding-variables
[12] https://explorable.com/conducting-an-experiment
[13] https://explorable.com/significance-test
[14] https://explorable.com/statistical-reliability
[15] https://explorable.com/validity-and-reliability
[16] https://explorable.com/types-of-validity
[17] https://explorable.com/instrument-reliability
[18] https://explorable.com/statistical-correlation
[19] https://explorable.com/test-retest-reliability
[20] https://explorable.com/cronbachs-alpha
[21] http://www.thefreedictionary.com/reliability