Methods of analysis and reliability

Test Validity and Reliability

Whenever a test or other measuring device is used as part of the data collection process, the validity and reliability of that test is important. Just as we would not use a math test to assess verbal skills, we would not want to use a measuring device for research that was not truly measuring what we purport it to measure. After all, we are relying on the results to show support or a lack of support for our theory and if the data collection methods are erroneous, the data we analyze will also be erroneous.

Test Validity. Validity refers to the degree in which our test or other measuring device is truly measuring what we intended it to measure. The test question "1 + 1 = _____" is certainly a valid basic addition question because it is truly measuring a student's ability to perform basic addition. It becomes less valid as a measurement of advanced addition because as it addresses some required knowledge for addition, it does not represent all of knowledge required for an advanced understanding of addition. On a test designed to measure knowledge of American History, this question becomes completely invalid. The ability to add two single digits has nothing do with history.

For many constructs, or variables that are artificial or difficult to measure, the concept of validity becomes more complex. Most of us agree that "1 + 1 = _____" would represent basic addition, but does this question also represent the construct of intelligence? Other constructs include motivation, depression, anger, and practically any human emotion or trait. If we have a difficult time defining the construct, we are going to have an even more difficult time measuring it. Construct validity is the term given to a test that measures a construct accurately and there are different types of construct validity that we should be concerned with. Three of these, concurrent validity, content validity, and predictive validity are discussed below.

Concurrent Validity. Concurrent Validity refers to a measurement device's ability to vary directly with a measure of the same construct or indirectly with a measure of an opposite construct. It allows you to show that your test is valid by comparing it with an already valid test. A new test of adult intelligence, for example, would have concurrent validity if it had a high positive correlation with the Wechsler Adult Intelligence Scale since the Wechsler is an accepted measure of the construct we call intelligence. An obvious concern relates to the validity of the test against which you are comparing your test. Some assumptions must

be made because there are many who argue the Wechsler scales, for example, are not good measures of intelligence.

Content Validity. Content validity is concerned with a test's ability to include or represent all of the content of a particular construct. The question "1 + 1 = ___" may be a valid basic addition question. Would it represent all of the content that makes up the study of mathematics? It may be included on a scale of intelligence, but does it represent all of intelligence? The answer to these questions is obviously no. To develop a valid test of intelligence, not only must there be questions on math, but also questions on verbal reasoning, analytical ability, and every other aspect of the construct we call intelligence. There is no easy way to determine content validity aside from expert opinion.

Predictive Validity. In order for a test to be a valid screening device for some future behavior, it must have predictive validity. The SAT is used by college screening committees as one way to predict college grades. The GMAT is used to predict success in business school. And the LSAT is used as a means to predict law school performance. The main

concern with these, and many other predictive measures is predictive validity because without it, they would be worthless.

We determine predictive validity by computing a correlational coefficient comparing SAT scores, for example, and college grades. If they are directly related, then we can make a prediction regarding college grades based on SAT score. We can show that students who score high on the SAT tend to receive high grades in college.

Test Reliability. Reliability is synonymous with the consistency of a test, survey, observation, or other measuring device. Imagine stepping on your bathroom scale and weighing 140 pounds only to find that your weight on the same scale changes to 180 pounds an hour later and 100 pounds an hour after that. Base don the inconsistency of this scale, any research relying on it would certainly be unreliable. Consider an important study on a new diet program that relies on your inconsistent or unreliable bathroom scale as the main way to collect information regarding weight change. Would you consider their results accurate?

A reliability coefficient is often the statistic of choice in determining the reliability of a test. This coefficient merely represents a correlation (discussed in chapter 8), which

measures the intensity and direction of a relationship between two or more variables.

Test-Retest Reliability. Test-Retest reliability refers to the test's consistency among different administrations. To determine the coefficient for this type of reliability, the same test is given to a group of subjects on at least two separate occasions. If the test is reliable, the scores that each student receives on the first administration should be similar to the scores on the second. We would expect the relationship between he first and second administration to be a high positive correlation.

One major concern with test-retest reliability is what has been termed the memory effect. This is especially true when the two administrations are close together in time. For example, imagine taking a short 10-question test on vocabulary and then ten minutes later being asked to complete the same test. Most of us will remember our responses and when we begin to answer again, we may just answer the way we did on the first test rather than reading through the questions carefully. This can create an

artificially high reliability coefficient as subjects respond from their memory rather than the test itself. When a pre-test and post-test for an experiment is the same, the memory effect can play a role in the results.

Parallel Forms Reliability. One way to assure that memory effects do not occur is to use a different pre- and posttest. In order for these two tests to be used in this manner, however, they must be parallel or equal in what they measure. To determine parallel forms reliability, a reliability coefficient is calculated on the scores of the two measures taken by the same group of subjects. Once again, we would expect a high and positive correlation is we are to say the two forms are parallel.

Inter-Rater Reliability. Whenever observations of behavior are used as data in research, we want to assure that these observations are reliable. One way to determine this is to have two or more observers rate the same subjects and then correlate their observations. If, for example, rater A observed a child act out aggressively eight times, we would want rater B to observe the same amount of aggressive acts. If rater B witnessed 16 aggressive acts, then we know at least one of these two raters is incorrect. If there ratings are positively correlated, however, we can be reasonably sure that they are measuring the same construct of aggression. It does not, however, assure that they are

measuring it correctly, only that they are both measuring it the same.

The test-retest reliability method is one of the simplest ways of testing the stability and reliability of an instrument over time.

For example, if a group of students takes a test, you would expect them to show very similar results if they take the same test a few months later. This definition relies upon there being no confounding factor during the intervening time interval.

Instruments such as IQ tests and surveys are prime candidates for test-retest methodology, because there is little chance of people experiencing a sudden jump in IQ or suddenly changing their opinions.

On the other hand, educational tests are often not suitable, because students will learn much more information over the intervening period and show better results in the second test.

Test-Retest Reliability and the Ravages of Time

For example, if a group of students take a geography test just before the end of semester and one when they return to school at the beginning of the next, the tests should produce broadly the same results.

If, on the other hand, the test and retest are taken at the beginning and at the end of the semester, it can be assumed that the intervening lessons will have improved the ability of the students. Thus, test-retest reliability will be compromised and other methods, such as split testing, are better.

Even if a test-retest reliability process is applied with no sign of intervening factors, there will always be some degree of error. There is a strong chance that subjects will remember some of the questions from the previous test and perform better.

Some subjects might just have had a bad day the first time around or they may not have taken the test seriously. For these reasons, students facing retakes of exams can expect to face different questions and a slightly tougher standard of marking to compensate.

Even in surveys, it is quite conceivable that there may be a big change in opinion. People may have been asked about their favorite type of bread. In the intervening period, if a bread company mounts a long and expansive advertising campaign, this is likely to influence opinion in favour of that brand. This will jeopardize the test-retest reliability and so the analysis that must be handled with caution.

Test-Retest Reliability and Confounding Factors

To give an element of quantification to the test-retest reliability, statistical tests factor this into the analysis and generate a number between zero and one, with 1 being a perfect correlation between the test and the retest.

Perfection is impossible and most researchers accept a lower level, either 0.7, 0.8 or 0.9, depending upon the particular field of research.

However, this cannot remove confounding factors completely, and a researcher must anticipate and address these during the research design to maintain test-retest reliability.

To dampen down the chances of a few subjects skewing the results, for whatever reason, the test for correlation is much more accurate with large subject groups, drowning out the extremes and providing a more accurate result.

Reproducibility is regarded as one of the foundations of the entire scientific method, a benchmark upon which the reliability of an experiment can be tested.

The basic principle is that, for any research program, an independent researcher should be able to replicate the experiment, under the same conditions, and achieve the same results.

This gives a good guide to whether there were any inherent flaws within the experiment and ensures that the researcher paid due diligence to the process of experimental design.

A replication study ensures that the researcher constructs a valid and reliable methodology and analysis.

Reproducibility vs. Repeatability

Reproducibility is different to repeatability, where the researchers repeat their experiment to test and verify their results.

Reproducibility is tested by a replication study, which must be completely independent and generate identical findings known as commensurate results. Ideally, the replication study should utilize slightly different instruments and approaches, to ensure that there was no equipment malfunction.

If a type of measuring device has a design flaw, then it is likely that this artefact will be apparent in all models.

The Process of Replicating Research

For most of the physical sciences, reproducibility is a simple process and it is easy to replicate methods and equipment.

An astronomer measuring the spectrum of a star notes down the instruments and methodology used, and an independent researcher should be able to achieve exactly the same results, Even in biochemistry, where naturally

variable living organisms are used, good research shows remarkably little variation.

However, the social sciences, ecology and environmental science are a much more difficult case. Organisms can show a huge amount of variation, making it difficult to replicate research exactly and so reproducibility is a process of attempting to make the experiment as reproducible as possible, ensuring that the researcher can defend their position.

In addition, these sciences have to make much more use of statistics to dampen down experimental noise caused by physiological and psychological differences between the subjects.

This is one of the reasons why most social sciences accept a 95% probability level, which is a contrast to the 99% confidence required by most physical sciences.

Reproducibility and Generalization - A Cautious Approach

Observing due caution with the process of generalizing results helps to strengthen the case for experimental reproducibility.

Generalization

In any study, there is a far smaller chance of finding confounding evidence if the claims are narrowly defined than if they are sweeping generalizations.

For example, a psychologist who found that aggression in children under the age of five increased if they watched violent TV, could generalize that all children under five would display the same condition.

Extending this to all children means that the experiment is prone to replication issues - A researcher finding that aggression did not increase in nine year old children would invalidate the entire premise by questioning the reproducibility.

Reproducibility is not Essential

It is important to understand that creating replicable research is not essential for validity, although it does help. Sometimes, due to sheer impracticality, temporal difficulties, or expense, this is not possible.

The Framingham Heart Study, an experiment testing three generations of nurses for cardiac issues, has been going on for over 60 years and nobody is seriously expected to replicate it.

Instead, results from other studies around the world are used to build up a database of statistical evidence supporting the findings.

Reproducibility - An Impossible Ideal?

Many scientists argue that reproducibility is not an important factor for many sciences observing natural phenomena, such as astronomy, geology and, notoriously, evolution.

The rise of the Intelligent Design movement has seen evolutionary science under attack, because creationists claim that evolution is not reproducible and, therefore, it is not valid. This has opened up an intense debate about the

role of replication study as, for example, a geologist cannot very well recreate conditions found on the primordial earth and observe rocks metamorphosing.

However, creationists misunderstand the idea of reproducibility and assume that it applies to an entire theory. In fact, this is incorrect and it is a manipulation of scientific practices; replicating research only applies to a specific experiment or observation.

Reproducibility and Specificity - A Geological Example

If I go into the Greek Mountains and observe trilobite fossils lying above ammonite fossils, I assume that trilobites came later than ammonites.

However, a more talented geologist than I later travels to exactly the same place, and points out that the rocks there are deformed and twisted 180 degrees, so my observations were the wrong way around. My field study was reproducible in that another researcher could come and try to replicate my observations.

Looking at the process from the other angle, imagine that an astronomer discovers a planet circling around a distant star. Nobody is suggesting that he builds a gaseous cloud and waits a few billion years for matter to accrete and an identical solar system to form, because that would be absurd.

Performing a replication study would involve other astronomers observing the star to try to find the planets, showing that there really are planets and that the original astronomer had no equipment malfunction.

Reproducibility and Archaeology - The Absurdity of Creationism

When Arthur Evans discovered Knossos, on Crete, and proposed that there was an ancient, advanced Minoan civilization, nobody suggested that he should recreate such a civilization and see if they built an identical city. Absurd as it may seem, this is the type of assumption that proponents of Creationism make.

Looking at this process in reverse, if a team of builders builds an exact replica of Knossos, it does not prove that such a civilization existed, although it would be a useful exercise in

looking at some of the techniques used by ancient builders, allowing archaeologists to refine their ideas. To suggest otherwise really is a deliberate misunderstanding and warping of the scientific method.

Ultimately, if Creationists use the argument that evolution is wrong because it is not reproducible, then they destroy their own argument. If evolutionary processes cannot be subjected to replicable research, neither can Intelligent Design, so their argument founders on its own presumptions. Surely, proponents of ID need to recreate the six days of Genesis before their ideas can be accepted by science!

Lecture: Methods of Structural Reliability Analysis

The aim of the present lecture is to introduce the most common techniques of structural

reliability analysis, namely, First Order Reliability Methods (FORM) and Monte-Carlo

simulation. First the concept of limit state equations and basic random variables is introduced.

Thereafter the problem of error propagation is considered and it is shown that FORM

provides a generalization of the classical solution to this problem. Different cases of limit

state functions and probabilistic characteristics of basic random variables are then introduced

with increasing generality. Furthermore, FORM results are related to partial safety factors

used in common design codes. Subsequently, crude Monte-Carlo and Importance sampling is

introduced as an alternative to FORM methods. The introduced methods of structural

reliability theory provide strong tools for the calculation of failure probabilities for individual

failure modes or components. On the basis of the present lecture, it is expected that the

students should acquire knowledge and skills in regard to:

- What is a basic random variable and what is a limit state function?
- What is the graphical interpretation of the reliability index?
- What is the principle for the linearization of non-linear limit state functions?
- How to transform non-normal distributed random variables into normal distributed

variables?

- How to consider dependent random variables?
- How are FORM results related to partial safety