

Quantitative Research: Reliability and Validity

Reliability

Definition: Reliability is the consistency of your measurement, or the degree to which an instrument measures the same way each time it is used under the same condition with the same subjects. In short, it is the repeatability of your measurement. A measure is considered reliable if a person's score on the same test given twice is similar. It is important to remember that reliability is not measured, it is estimated.

There are two ways that reliability is usually estimated: test/retest and internal consistency.

Test/Retest: Test/retest is the more conservative method to estimate reliability. Simply put, the idea behind test/retest is that you should get the same score on test 1 as you do on test 2. The three main components to this method are as follows:

- 1.) implement your measurement instrument at two separate times for each subject;
- 2.) compute the correlation between the two separate measurements; and
- 3) assume there is no change in the underlying condition (or trait you are trying to measure) between test 1 and test 2.

Internal Consistency

Internal consistency estimates reliability by grouping questions in a questionnaire that measure the same concept. For example, you could write two sets of three questions that measure the same concept (say class participation) and after collecting the responses, run a correlation between those two groups of three questions to determine if your instrument is reliably measuring that concept.

One common way of computing correlation values among the questions on your instruments is by using Cronbach's Alpha. In short, Cronbach's alpha splits all the questions on your instrument every possible way and computes correlation values for them all (we use a computer program for this part). In the end, your computer output generates one number for Cronbach's alpha - and just like a correlation coefficient, the closer it is to one, the higher the reliability estimate of your instrument. Cronbach's alpha is a less conservative estimate of reliability than test/retest.

The primary difference between test/retest and internal consistency estimates of reliability is that test/retest involves two administrations of the measurement instrument, whereas the internal consistency method involves only one administration of that instrument.

Validity

Definition: Validity is the strength of our conclusions, inferences or propositions. More formally, Cook and Campbell (1979) define it as the "best available approximation to the truth or falsity of a given

inference, proposition or conclusion." In short, were we right? Let's look at a simple example. Say we are studying the effect of strict attendance policies on class participation. In our case, we saw that class participation did increase after the policy was established. Each type of validity would highlight a different aspect of the relationship between our treatment (strict attendance policy) and our observed outcome (increased class participation).

Types of Validity:

There are four types of validity commonly examined in social research.

1. Conclusion validity asks if there is a relationship between the program and the observed outcome? Or, in our example, is there a connection between the attendance policy and the increased participation we saw?
2. Internal Validity asks if there is a relationship between the program and the outcome we saw, is it a causal relationship? For example, did the attendance policy cause class participation to increase?
3. Construct validity is the hardest to understand in my opinion. It asks if there is a relationship between how I operationalized my concepts in this study to the actual causal relationship I'm trying to study? Or in our example, did our treatment (attendance policy) reflect the construct of attendance, and did our measured outcome - increased class participation - reflect the construct of participation? Overall, we are trying to generalize our conceptualized treatment and outcomes to broader constructs of the same concepts.
4. External validity refers to our ability to generalize the results of our study to other settings. In our example, could we generalize our results to other classrooms?

Threats to Internal Validity

There are three main types of threats to internal validity - single group, multiple group and social interaction threats.

Single Group Threats apply when you are studying a single group receiving a program or treatment. Thus, all of these threats can be greatly reduced by adding a control group that is comparable to your program group to your study.

A **History Threat** occurs when an historical event affects your program group such that it causes the outcome you observe (rather than your treatment being the cause). In our earlier example, this would mean that the stricter attendance policy did not cause an increase in class participation, but rather, the expulsion of several students due to low participation from school impacted your program group such that they increased their participation as a result.

A **Maturation Threat** to internal validity occurs when standard events over the course of time cause your outcome. For example, if by chance, the students who participated in your study on class participation all "grew up" naturally and realized that class participation increased their learning (how likely is that?) - that could be the cause of your increased participation, not the stricter attendance policy.

A **Testing Threat** to internal validity is simply when the act of taking a pre-test affects how that group does on the post-test. For example, if in your study of class participation, you measured class

participation prior to implementing your new attendance policy, and students became forewarned that there was about to be an emphasis on participation, they may increase it simply as a result of involvement in the pretest measure - and thus, your outcome could be a result of a testing threat - not your treatment.

An *Instrumentation Threat* to internal validity could occur if the effect of increased participation could be due to the way in which that pretest was implemented.

A *Mortality Threat* to internal validity occurs when subjects drop out of your study, and this leads to an inflated measure of your effect. For example, if as a result of a stricter attendance policy, most students drop out of a class, leaving only those more serious students in the class (those who would participate at a high level naturally) - this could mean your effect is overestimated and suffering from a mortality threat.

The last single group threat to internal validity is a *Regression Threat*. This is the most intimidating of them all (just its name alone makes one panic). Don't panic. Simply put, a regression threat means that there is a tendency for the sample (those students you study for example) to score close to the average (or mean) of a larger population from the pretest to the posttest. This is a common occurrence, and will happen between almost any two variables that you take two measures of. Because it is common, it is easily remedied through either the inclusion of a control group or through a carefully designed research plan (this is discussed later). For a great discussion of regression threats, go to [Bill Trochim's Center for Social Research Methods](http://www.socialresearchmethods.net/tutorial/Colosi/lcolosi2.htm).

In sum, these single group threats must be addressed in your research for it to remain credible. One primary way to accomplish this is to include a control group comparable to your program group. This however, does not solve all our problems, as I'll now highlight the multiple group threats to internal validity.

Multiple Group Threats to internal validity involve the comparability of the two groups in your study, and whether or not any other factor other than your treatment causes the outcome. They also (conveniently) mirror the single group threats to internal validity.

A *Selection-History* threat occurs when an event occurring between the pre and post test affects the two groups differently.

A *Selection-Maturation* threat occurs when there are different rates of growth between the two groups between the pre and post test.

Selection-Testing threat is the result of the different effect from taking tests between the two groups.

A *Selection-Instrumentation* threat occurs when the test implementation affects the groups differently between the pre and post test.

A *Selection-Mortality* Threat occurs when there are different rates of dropout between the groups which lead to you detecting an effect that may not actually occur.

Finally, a *Selection-Regression* threat occurs when the two groups regress towards the mean at different rates.

Okay, so know that you have dragged yourself through these extensive lists of threats to validity - you're wondering how to make sense of it all. How do we minimize these threats without going insane in the process? The best advice I've been given is to use two groups when possible, and if you do, make sure they are as comparable as is humanly possible. Whether you conduct a randomized experiment or a non-random study --> YOUR GROUPS MUST BE AS EQUIVALENT AS POSSIBLE! This is the best way to strengthen the internal validity of your research. The last type of threat to discuss involves the social pressures in the research context that can impact your results. These are known as **social interaction threats** to internal validity.

Diffusion or "Imitation of Treatment" occurs when the comparison group learns about the program group and imitates them, which will lead to an equalization of outcomes between the groups (you will not see an effect as easily).

Compensatory Rivalry means that the comparison group develops a competitive attitude towards the program group, and this also makes it harder to detect an effect due to your treatment rather than the comparison group's reaction to the program group.

Resentful Demoralization is a threat to internal validity that exaggerates the posttest differences between the two groups. This is because the comparison group (upon learning of the program group) gets discouraged and no longer tries to achieve on their own.

Compensatory Equalization of Treatment is the only threat that is a result of the actions of the research staff - it occurs when the staff begins to compensate the comparison group to be "fair" in their opinion, and this leads to an equalization between the groups and makes it harder to detect an effect due to your program.

Threats to Construct Validity

I know, I know - you're thinking - no I just can't go on. Let's take a deep breath and I'll remind you what construct validity is, and then we'll look at the threats to it one at a time. OK? OK.

Construct validity is the degree to which inferences we have made from our study can be generalized to the concepts underlying our program in the first place. For example, if we are measuring self-esteem as an outcome, can our definition (operationalization) of that term in our study be generalized to the rest of the world's concept of self-esteem?

Ok, let's address the threats to construct validity slowly - don't be intimidated by their lengthy academic names - I'll provide an English translation.

Inadequate Preoperational Explication of Constructs simply means we did not define our concepts very well before we measured them or implemented our treatment. The solution? Define your concepts well before proceeding to the measurement phase of your study.

Mono-operation bias simply means we only used one version of our independent variable (our program or treatment) in our study, and hence, limit the breadth of our study's results. The solution? Try to implement multiple versions of your program to increase your study's utility.

Mono-method bias simply put, means that you only used one measure or observation of an important concept, which in the end, reduces the evidence that your measure is a valid one. The solution? Implement multiple measures of key concepts and do pilot studies to try to demonstrate that your measures are valid.

Interaction of Testing and Treatment occurs when the testing in combination with the treatment

produces an effect. Thus you have inadequately defined your "treatment," as testing becomes part of it due to its influence on the outcome. The solution? Label your treatment accurately.

Interaction of Different Treatments means that it was a combination of our treatment and other things that brought about the effect. For example, if you were studying the ability of Tylenol to reduce headaches and in

actuality it was a combination of Tylenol and Advil or Tylenol and exercise that reduced headaches -
- you would have an interaction of different treatments threatening your construct validity.

Restricted Generalizability across Constructs simply put, means that there were some unanticipated effects from your program that may make it difficult to say your program was effective.

Confounding Constructs occurs when you are unable to detect an effect from your program because you may have mislabeled your constructs or because the level of your treatment wasn't enough to cause an effect.

As with internal validity, there are a few social threats to construct validity also. These include:

1. **Hypothesis Guessing:** when participants base their behavior on what they think your study is about - so your outcome is really not due solely to the program - but also to the participants' reaction to you and your study.
2. **Evaluator Apprehension:** When participants are fearful of your study to the point that it influences the treatment effect you detect.
3. **Experimenter Expectancies:** when researcher reactions shape the participant's responses - so you mislabel the treatment effect you see as due to the program when it is more likely due to the researcher's behavior.

See, that wasn't so bad. We broke things down and attacked them one at a time. You may be wondering why I haven't given you a long list of threats to conclusion and external validity - the simple answer is it seems as if the more critical threats involve internal and construct validity. And, the means by which we improve conclusion and external validity will be highlighted in the section on [Strengthening Your Analysis](#).

Summary

The real difference between reliability and validity is mostly a matter of definition. Reliability estimates the consistency of your measurement, or more simply the degree to which an instrument measures the same way each time it is used in under the same conditions with the same subjects. Validity, on the other hand, involves the degree to which you are measuring what you are supposed to, more simply, the accuracy of your measurement. It is my belief that validity is more important than reliability because if an instrument does not accurately measure what it is supposed to, there is no reason to use it even if it measures consistently (reliably).