# Validity and Reliability: Quantitative Research Study Buddies

Good quantitative research studies are designed to be as valid and reliable as possible, while recognizing the constraints that we have when doing research in the real world.[1] What does this mean? What enhances the validity and reliability of a research study?

## ✸ **Validity** – are we measuring what we purport to measure?

The validity of a study depends on the design of the research setup all the way through to the interpretation of the results. It starts with the instrument or measurement system you are using to collect data that is relevant to your research question. An instrument could be a grading system, a survey, a methodology for collecting observations, a test, or some other measure of a phenomena. To support validity, we ask ourselves:

1. Is the instrument or assessment a good measurement tool? Is it aligned with the research question?
2. During the study, are there internal or external factors that impact validity?
3. And then, is your interpretation of the results aligned with what you measured and how you measured it?

**The first question** is answered by ensuring that you either use an instrument or measurement system that has been tested by others, or that you test it yourself. For example, if you are trying to test students' ability to apply basic concepts in physics, you might use the Force Concept Inventory (FCI), which is a standard exam that has been extensively tested/validated.[2] Alternatively, you could create your own exam or other measurement system, but then you would want to go through a process of developing an instrument based on theory or findings from the literature, and testing and refining your proposed tool before using it in your study.

The corollary to this: Is your measurement system aligned with your research question? You can have a highly validated assessment instrument (e.g., the FCI), but if you misuse it then it is not a valid measure (e.g., use the FCI to measure advanced physics topics). Examine your research question carefully. Are you using a measurement approach that is as direct and aligned as possible with the question that you are trying to address?

**The second question** is answered during your study. Did something happen to threaten validity? This could include an unusual pattern of people dropping out of your study, or perhaps your study design did not account for differences between groups (Creswell has a couple of nice tables showing types of threats to validity, p. 169-172).

**The third question**: Are you interpreting your results in a manner that aligns with what the instrument purports to measure and the way you have used it. That is, are the claims you are making in your discussion and conclusions based on valid data? Do your claims follow from, and align with, what you measured? Heale and Twycross (2015), for example, note that you cannot draw conclusions about depression in a population using a survey that measures anxiety. The resulting data are not valid for this purpose.

## ✸ **Reliability** – are our measurements dependable/consistent?

The reliability of a study design substantially depends on the method and actual circumstances happening when the study was running. That is, if you ran the study again, in the same context with comparable participants, would you get the same results? Are your results dependable? Reliability is enhanced, for example, by having a larger sample size, a sample that is representative of your population of interest, and ensuring that the circumstances during the trial or observation did not substantially interfere with the study.

---

[1] In qualitative studies some authors (Creswell) talk about validity. However, others use the analogous concept of *methodological integrity* which includes *fidelity* and *utility* (see *Reporting Qualitative Research in Psychology*, H.M. Levitt, 2020).

[2] The Force Concept Inventory was originally proposed in 1985 by Hestenes, Wells, and Swackhamer. There is an extensive literature on this standard test.

Let's consider the FCI again. Suppose you give a group of students the FCI before they take a physics course and at the end. This is called a pretest-posttest study, and you are looking for gain, that is; the degree to which students demonstrate better understanding of basic physics at the end of the course. But suppose the power in the building goes out during the posttest so the students are writing the test in the dark, or suppose you only test 3 students in the class, or suppose your posttest is at 8:00am the day after the prom. All these circumstances (some under your control and some not) will erode reliability – to a greater or lesser extent. Just as in science experiments, running more trials, making sure external factors don't mess with your measurements, and so on, improve reliability. Researchers use various reliability coefficients (which essentially represent correlations) to inspect reliability. One example is Cronbach's α, commonly used for surveys. There are others used for tests and other methodologies.[3]

## ✸ **Limitations** – all studies have them

Is any study perfectly valid and reliable? Absolutely not. There is just no such thing. The reality of research, especially in engineering education, is that you usually cannot run a survey or test an intervention many, many times over years and years. And you can only validate your measurement approach so many times in so many ways given resource and time limitations. So how do you effectively manage validity and reliability?

First, pay close attention to your study design. Have you chosen a measurement approach that has strong validity for what you are trying to measure? If you are creating a measurement (e.g., a survey, rubric, test, etc.) then also design a testing regime for your measurement system before deploying it. Think ahead to reporting your results: how will you explain the alignment with your research question? Plan out your method thoughtfully: how will you deploy this measurement? Think about the ways you can improve the reliability of your results, and report what you did to improve reliability and any evidence that can demonstrate their reliability. Not all aspects of reliability are under your control, but the design of your study methods can improve the chances of enhancing reliability.

Next, check the claims you are making about the results. Many people think about validity in terms of the results: are the data valid given the design of the study, and what happened during the study? The validity of the results is strengthened when the research question and measurement method are well-aligned; an appropriate, validated instrument (measurement approach) is used; you managed internal and external threats to validity during the study, and the interpretation of the results is aligned with what the instrument measures.

Then, use the limitations section of your write-up (in your thesis, or paper) to explain the steps you took to enhance reliability and validity and the limitations in the study design and deployment. Maybe a pandemic hit in the middle of your study. Maybe you had to use convenience sampling because it was the only feasible option. Also clearly explain the limitations of your results and the generalizability of your conclusions. Readers understand that real-life research has real limitations, and they want to hear how these limitations impact the data and interpretation of your results.

One common question: I know there are limitations in my study, but I'm not sure whether to call it an issue of validity or reliability. How do I explain this? Answer: sometimes it is a matter of interpretation whether an issue is one of validity or reliability. Instead of trying to label it, consider simply explaining the issue, and pointing out how you accounted for it in your results/discussion, or how it may impact interpretation or generalizability of your work.

**Author: Susan McCahan, University of Toronto**
Acknowledgement: Thank you also to Brian Frank (Queen's U.), Qin Liu (U. Toronto), Sean Maw (U. Saskatchewan), and Cindy Rottmann (U. Toronto) for contributing to this snack.

**References:**

J.W. Creswell and J.D. Creswell, Research Design: Qualitative, Quantitative and Mixed Methods Approaches, 5th ed., Sage, 2018.
A. Field, "Discovering Statistics using IBM SPSS Statistics", 4th ed., Sage, 2013.
R. Heale and A. Twycross, "Validity and reliability in quantitative studies," Evid Based Nurs, Vol. 18(3), 2015.
B.M. Moskal, J.A. Leydens, M.J Pavelich, "Validity, Reliability and the Assessment of Engineering Education", JEE, 2002.



[3] https://www.ets.org/Media/Research/pdf/RM-18-01.pdf