Some Basic Threats to Experimental Validity

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Threats to Validity

1. Introduction
2. Threats to Internal Validity
   - Selection
   - Selection by Maturation Interaction
   - Regression Artifacts
   - Experimenter Bias
   - Maturation
   - History
   - Mortality
   - Instrumentation
3. Threats to External Validity
   - Demand Characteristics
   - Interaction between Selection and the Experimental Variable
   - The File Drawer Problem

Reliability and Validity

- At the start of the course, I mentioned that good statistics cannot rescue bad data
- One way that people generate bad data in their reseach is through measures that have low validity or low reliability.

Reliability and Validity

Reliability

- Assume that a measure X is trying to measure a construct Y.
- The measure has high *reliability* if it repeatedly generates the same result when it is measuring the same value of the construct.
- Reliable measurements are replicable and have "low noise."
- Later in the course, we will discuss ways in which we can measure the reliability of a test or measure.

Reliability and Validity

Validity

- A measure is said to have high *construct validity* if it measures what it is supposed to measure.
- A weaker form of validity is *face validity*. A measure has high *face validity* if it appears, on the basis of observable characteristics, to have a reasonable likelihood of measuring what it is supposed to measure.

Threats to Experimental Validity

- Basic experiments attempt to manipulate an independent variable while holding all other factors constant, and observe the effect on a dependent variable.
- Although this notion is simple in concept, it is very difficult to execute in practice.
- Many factors threaten the validity of even the simplest experimental design.
- In this lecture, we'll review some of the more basic threats to experimental validity.
- As basic as they are, they are very pervasive in modern research.

Internal and External Validity

- Suppose we manipulate X with the intention of determining whether it affects Y.
- The experiment has *internal validity* if, *within the confines of the experiment*, it may be reliably concluded whether X affected Y.
- The experiment has *external validity* if its findings about causality generalize beyond the specific experimental setting and studied sample to more “the world at large.”
- We will now examine some of the most basic threats to internal and external validity.
- We’ll start with internal validity in the next section.

Selection
Self-Selection

- The *selection* problem occurs when groups are not equivalent because participants have somehow self-selected or have been selected in non-random fashion.
- *Self-selection*, in which subjects decide which experimental group they will be in, will often ruin a study.
- Often the problem is exceedingly obvious.

Example (A Marijuana Study)

Suppose the experimenter posted a signup sheet saying that the Marijuana Group would smoke two large marijuana cigarettes and the Control Group would drink two cups of coffee prior to a complex cognitive test, and that subjects should sign up for the group that they wanted to be in. Describe some of the possible selection effects.

Selection

Faulty “Random” Assignment

- Sometimes ”random” assignment is really not random at all.
- For example, splitting a room down the middle and assigning people on the left to one group and people on the right to another *might* seem reasonable.
- But is it?

Example (Faulty Random Assignment)

One year, I discovered that there was a substantial difference in performance between those who sat on the right and those who sat on the left in my large undergraduate statistics lecture. What might have caused this?

Selection by Maturation Interaction

- Some groups will naturally grow apart as they mature.
- These changes can be interpreted incorrectly as effects of an experimental manipulation.
- The problem is especially prevalent when studying naturally intact groups.

Selection by Maturation Interaction

Example (Selection by Maturation Interaction)

Suppose a study is run to determine the effects of Vitamin S on Strength development. The experimenters take two intact classes of 6th graders, administer Vitamin S to Group I, and a placebo to Group II over a three year period, then measure them again in 9th grade.

Results. In 6th grade, the two groups had virtually identical averages on a test of strength. When tested again in 9th grade, the groups had grown apart. Group I was substantially stronger on average.

Conclusion. The initial conclusion was that Vitamin S had caused an increase in strength in Group I.

Follow-Up. On closer examination, it was found that Group I was 67% male, 33% female, while Group II was 47% male and 53% female. The greater increase in strength was due to the disparity in the number of males in the two groups.

Regression Artifacts

- When measures are not totally reliable, a portion of the score that is obtained is a random component that might be considered, informally, as a "luck factor."
- For example, a course exam is not a perfect indicator of your knowledge of the course material. Part of your performance is due to luck.
- Can you identify several aspects of "luck" that contribute to your performance and might be considered random?

Regression Artifacts

- We might say that $X = T + E$, your exam score is composed of a “true score component” and a “random error component.”
- Suppose I give the first exam in the course, and I select the people with the five highest grades in the class.
- All other things being equal, would you expect these 5 people to have had a positive or a negative E (luck component) on the exam?
- So, all other things remaining equal, what would you expect their performance on the second exam to be, relative to their performance on the first exam?

Regression Artifacts

- Now, suppose I selected the 5 students who had the lowest marks in the class.
- What about *their* luck component?
- All other things being equal, what would we expect to happen to their performance on Exam 2 relative to Exam 1?

Regression Artifacts

Example (Regression Effects)

In the early research on Early Childhood Enrichment, some researchers did not control for regression effects. They selected children who had scored extremely low on standardized IQ tests, and put them in special enrichment programs. They showed dramatic improvement. Unfortunately a substantial amount of the improvement was a regression artifact.

Such effects can be controlled for by including a no-treatment or waiting list control.

Experimenter Bias

- If the experimenter knows what group a subject is in, then there is a chance that the experimenter will behave differently toward that subject and influence the subject's behavior in a way that changes the outcome of the study.
- This can occur with no conscious effort on the experimenter's part.
- The difference can be extremely subtle, but the impact on the study can be very large.

Experimenter Bias

- The typical way of protecting against experimenter bias is to use randomization with *Double Blind Controls*, in which the subject does not know what group he/she is in, *and* the experimenter does not know what group the subject is in.
- Some famous studies have not controlled for experimenter bias.

Maturation

- Maturation in this context is a technical term used to refer to changes that occur as a result of processes within participants as a function of time
- For example,
  1. Aging in studies that occur over long periods of time
  2. Participants getting tired and hungry in studies that occur over several hours

History

- History refers to specific events occurring between pre-test and post-test that are external to participants, independent of experimental manipulation, and have an effect on post-test results.
- An example: A study on the effect of several kinds of persuasive communication regarding a political candidate would be disrupted if a scandal erupted regarding the candidate's personal life between the pre-test and post-test.

Mortality

- In the context of experimental design, this term refers to any factor that causes subjects to drop out of a study.
- Differential drop-out rates across groups produce spurious differences between groups.
- It is ofteh impossible to determine what caused different drop-out rates and how they affected results.

Instrumentation

- This term is used very broadly to refer to any systematic changes in the instruments, people or procedures used to produce the data in the experiment.
- Some examples include:
  1. A scale goes out of calibration.
  2. A rater gets sick and is replaced by another rater with different standards.
  3. A questionnaire is administered several times in a longitudinal study. The experimenter runs out of copies of the questionnaire, and, unknown to her, the new copies of the questionnaire have some revised items.

Demand Characteristics

- The subject in an experiment usually knows he/she is in an experiment.
- In many cases, the various manipulations become pretty transparent, for a number of reasons.
- In such cases, the subject may come to realize that the experimenter is expecting a certain kind of behavior.
- Depending on the personality characteristics of subject and experimenter, the subject may respond in a way that is not typical of the way subjects "in the real world" would respond.
- In such cases, the experiment may have internal replicability and be internally valid, but have no serious implications for the way people respond in the real world.

Selection and the Experimental Variable

- In some cases, the nature of the experimental variable itself interacts with the availability of subject populations.
- For example, suppose your advisor is trying to recruit subjects to participate in an avant garde program to provide sex education for kindergarten students.
- The nature of the subject matter being studied may impact on the kind of school district that will allow you access to their students to do research.
- As a result, your study may not generalize to all school populations.

The File Drawer Problem

- We often assume that published academic research, or research produced by professionals in an industrial setting, is “representative” in the sense that it is unbiased (although may be subject to “the luck of the draw”)
- Consequently, if two or more studies find the same result, the tendency is to believe that the result is a valid representation of reality
- However, this need not be so because of *The File Drawer Problem.*

The File Drawer Problem

- In many fields of research, there is a strong bias toward publishing only *statistically significant results*, that is, results in which the independent variable was found to affect the dependent variable.
- So experimenters who fail to get significant results just file their articles away, rather than submit them for publication.
- Moreover, for reasons that will become clearer later in the course, experimenters who submit non-significant results for consideration for publication often get them rejected.
- So suppose 10 researchers run experiments on the same idea, and only two get statistically significant results. Because of the file drawer problem, the two significant results may be the only ones that ever see "the light of day."

The File Drawer Scam

- A variation of the file drawer problem is used to trick unwary consumers.

Example (Incredibly Accurate Stock Picks)

Several years ago, I received a junk mail ad from a "financial advising service" that claimed a very high success rate at picking stocks. The letter listed 4 stocks that it rated as "best buys." I read it, and without thinking dropped it in a corner of my desk, where it was soon buried in a pile of other items. A half year later, I received a second letter from the same company. It started by saying "Six months ago, we offered you a chance to subscribe to our newsletter at a discount rate. If you had bought our 4 picks that week, you would, by now, have doubled your original investment."

A short while later, I found the original letter on my desk. Sure enough, if I had purchased the 4 stocks they touted, I would have doubled my investment in six months. Wow!

Is it possible that they tricked me? How?