Chapter 10: Evaluating Causality with Experiments

Experiments vs. Observational Studies

When determining causality, researchers may use either an observational study or an experiment to collect the appropriate data.

Observational studies can only observe what explanatory and response outcomes are associated, but without assigning units to particular interventions. We leave open the possibility that the units engaged in each intervention may have systematic differences that affect the response!

In contrast, **experiments** directly assign units to an intervention to see if response outcomes are directly affected. The goal is to isolate some factor as the only explanation for a change in response.

Investigation: On the Netflix show *100 Humans*, the researchers wanted to know—Can good looks keep you out of jail? Let's identify some experimental features.

Unit of observation: 1 (adult, U.S.) "human"

Population: All (adult U.S.) humans

Explanatory Variable:

Which we may further break down into…

Treatment factor:

Control factor:

Response variable:

This study was likely **blinded**. That means…

Double Blinding means that the people administering the intervention also do not know who is in which group. Would you guess that this study was double blinded?

Some studies may use a **Placebo—**a non-effective substance/intervention that is designed to mimic the interventional _____ of the treatment factor. This study didn't use a placebo in the strictest sense, but did use a carefully chosen comparative intervention.

Good experiments identify differences in the response that can *only* be attributed to the **treatment factor** and *nothing else*! They should do a good job eliminating possible confounders to the causal link…but not all experiments succeed in doing that.

Single group Pre-Post Designs: All units complete the same intervention(s) in the same order. We then compare the pre and post measures to see if there is a systematic difference on average.

Control (Intervention) → Pre-measure → Treatment Intervention → Post-measure

- **Investigation (Sleep Aid Study):** Developers of a new sleep aid study its effect on improving average duration of sleep. To study this, the researchers select 100 people who report issues with sleeping.
    - First, participants report their nightly sleep amount and quality for 2 weeks prior to using any sleep aid.
    - Second, participants are given a 2 week supply of the sleep aid and asked to take it before bed. They again report their nightly sleep amount and quality for 2 weeks while on the sleep aid.

  The researchers noticed that sleep levels and sleep quality was higher on average during the 2 week period that participants took the sleep aid. Does that suggest the sleep aid directly increased sleep level/quality? Are there any other explanations for this difference besides the sleep aid?

- **Investigation (Reward vs. Punishment):** Another *100 Humans* experiment examined: Under what conditions humans perform better

  What else *might* explain the difference in response values observed here? Is the instruction type the only systematic difference?

To summarize, pre-post studies should be used cautiously due to confounding threats from…

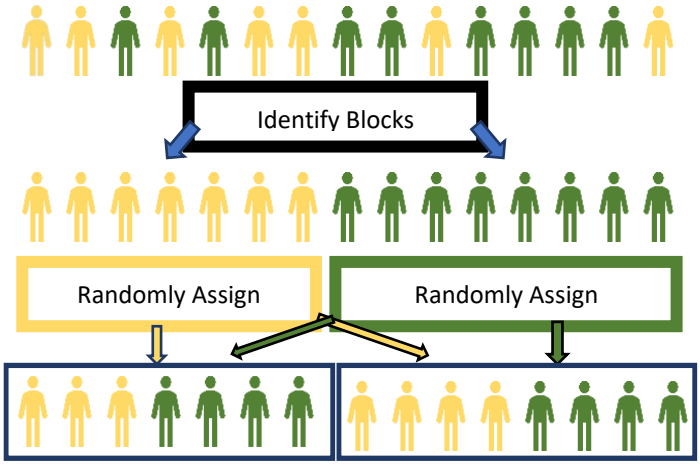
Multi-group designs: In a multi-group design, we can now separate the treatment/control factors into separate groups and potentially avoid other confounding differences, such as timing differences, test familiarity, or reactance/placebo effects. There are *several* design types and features we'll discuss.

Treatment Intervention → Response measure

Control (Intervention) → Response measure

But now that we have two groups, we need to be confident that these two groups are _____ and have no systematic differences between them.

- **Randomized Controlled Experiments** use random assignment to sort units—it may be pure random assignment, or random assignment with blocking by some relevant factors.
  - ***Random Assignment*** means using a random chance process to sort units.
  - ***Random Assignment with Blocking*** first involves blocking units by possible _____ factors, and then randomly assigning from each block.
    - Blocking is identifying individual characteristics (like age, medical condition, sex, etc.) that might interact with the treatment or affect the response.
    - Then after blocking, the researchers randomly assigns units in each block to a group.
    - Pure random assignment works very well with larger groups (e.g., n> 50), but blocking can help ensure equivalent groups when groups are _____.



- **Non-randomized Controlled Experiment**: Uses a non-random assignment method to sort groups.
  - Be cautious with non-randomized sorting methods. They may create a systematic bias in our groups!

For each description below, identify the sorting method used.

Sorting by last name alphabetically. First half of alphabet to one group, the rest to the other group.

Sorting by coin flips. "Heads" goes to one group. Once half have been assigned to a group, remainder go to other group.

Using Excel random number generator to sort half of 18-35 year olds, half of the 36 – 65 and half of 65+ to one group. The rest to the other group.

Additional Multi-group features: There are some nice features to single group pre-post designs that are missing from the basic multi-group template we outlined above.

1. In a single group pre-post design, we record the before and after data for each participant. Having a before and after measure for each participant might be advantageous to have.

- **Randomized Controlled Experiment with *Repeated measures***
    - By taking multiple measures per person at different time points, we can better track how individuals respond to the treatment and understand individual variation.
    - Note that it may not *always* be advantageous to take multiple measurements depending on what type of response measurements you are taking—like large test familiarity threats.

Response Measure → Treatment Intervention → Response measure

Response Measure → Control Intervention → Response measure

2. In a single group pre-post design, we can get treatment factor responses from all our participants, rather than only half. This also avoids the ethical dilemma of only assigning some participants to the more effective treatment condition.

- **Randomized Controlled Experiment with *Crossover trials***
    - Crossover trials are a *special case* of repeated measures. Each group completes both experimental conditions and produce response measures from both.

Treatment Intervention → Response meas. → Control (Intervention) → Response meas.

Control (Intervention) → Response meas. → Treatment Intervention → Response meas.

    - In crossover trials, researchers do need to be wary of **lingering effects** during the second round. So in some studies, researchers may add _____ between each phase.

Investigation Reconsidered: How might the Reward vs. Punishment experiment on 100 Humans be redesigned?

Reflection Questions

10.1. Consider an experiment to determine whether a particular vaccine is effective at lowering the risk of infection from HPV. What is the explanatory variable and what is the response variable? What is the treatment factor?

10.2. What does it mean if a study is blinded? What does it mean if it's double blinded? Does having a placebo that is identical to the treatment factor make it easier to blind a study or harder?

10.3. Which causality threats did we identify as common to single group pre-post designs?

10.4. How is random assignment with blocking different than pure random assignment? In what situation is pre-assignment blocking especially critical in ensuring equivalent groups?

10.5. How is a randomized controlled experiment with repeated measures different than simply a randomized controlled experiment? What value might repeated measures add to the design?

10.6. What is a randomized controlled experiment with cross-over trials? What value might cross-over trials add to the design? What threat should we be wary of when doing cross-over trials?

Investigation (Mathbar): A large, randomized controlled experiment assessed if students' mathematical performance was enhanced by taking *MathBar*—a specially-formulated protein bar. Half of the participants were given *MathBar* before the test and were told this would boost their focus and memory recall. They completed their exam in one classroom. The other half did not receive anything and served as a control group. They completed the same exam in another classroom. The group that received *MathBar* had a "statistically significant" higher average score. The researchers claimed: "*MathBar* improves students' mathematical performance."

Does this study provide good evidence that MathBar caused an increase in mathematical performance?

- **Threats to Causality Summarized** *(not an exhaustive list, but a good start!)*
  - **Group Selection** – Are there any systematic differences between our groups?
    - If we are comparing two or more groups in an experiment, the groups should be similar.
    - Random assignment with large groups, or random assignment with blocking for smaller groups are the best way to guard against systematic differences between groups.

  - **Drop Out Differences** – Did drop-out differences introduce non-equivalency at the end?
    - _____ is a term for when participants do not continue in the study. Perhaps we lose contact with them, or perhaps they don't adhere fully to their treatment plan.
    - _____ is a term for when participants might pass away during the study, perhaps as a result of the condition being treated or the treatment itself.
    - Some attrition or mortality is expected in medical studies. But drop out differences can threaten the causality argument when 1) drop out _____ are different between group or when 2) drop out _____ are systematically different between each group.
    - In cases where there is some level of attrition due to non-adherence, researchers may make a comparison of all _____ participants to ensure balance.

  - **Test Familiarity** – Are participants simply getting better at completing the measure?
    - This threat would be most pertinent when the instrumentation is a duplicated mental or physical test. Participants are getting an opportunity to practice or learn from the test!
    - This problem is exacerbated in a single group _____ when there is no control group to compare that test familiarity bump to.
    - Test familiarity typically _____ in multi-group designs, but test familiarity could still weaken the validity of the instrument itself as a reliable post-measure.

- **Timing Effects** – Do systematic differences in group timing affect outcomes?
  - If there are systematic differences between the groups' intervention times, that could lead to timing-related confounders (e.g., current events, weather, time-of-day differences).
  - *Note: It is **ok** if **individuals** happen to complete their intervention or response measures at **different times**—the question is whether one group is systematically earlier/later than another!*

- **Setting Effects** – Do any other setting or experiential differences affect the response?

  - **Placebo Effect*:*** Are participants improving just because they know they are receiving something. *This is a concern when we _____ have an appropriate placebo/comparison treatment for the control group, or no comparison group at all.*
  - **Researcher Effects*:*** If researchers interacting with the participants know who is in which group, they *may* act differently around each group. Use _____ when this is a significant threat.
  - **Environment Condition Differences*:*** Are environmental conditions different between the treatment and control conditions beyond the treatment factor you wish to study? People differences? Location differences? Context differences? *As with timing differences, the concern is group differences, not individual differences!*

- **Independence** – Are the units in each group providing independent response outcomes?
  - In experiments where people might interact with one another in their group, group dynamics may threaten the independence of our data.
  - In extreme cases, group dynamics could turn your group of, say 30 people, into a monolith, resulting in a functional comparison of sample sizes of __.

In general…Ask whether the treatment factor has ***clearly** been **isolated*** in the comparison. Choosing an appropriate placebo or comparative intervention is important!

Investigation (Gender Bias): Let's look at one more clip on Gender Bias from *100 Humans*.

Reflection Questions

10.7. Can you name the 6 categories of causality threats we learned in this chapter?

10.8. Consider the HPV study outlined in the first reflection question. Which category do you think poses more of a causality threat to a study of this nature—a test familiarity threat or a drop-out difference threat?

10.9. In the 100 Humans clip on perceptions of gender differences, did the researchers do a good job isolating the causal effect of gender on human perception? Why or why not?

Chapter 10 Additional Practice (Videos available in the Ch 10 module on Canvas!)

Practice: A study investigates if Gatorade truly improves endurance in cardio-intensive sports. In a study of 100 athletes, 50 were assigned to drink Gatorade while 50 were assigned to drink Water. The athletes were then asked to cycle at a certain speed for as long as they could. The research team recorded how long each participant kept their pace.

The unit of observation:

the treatment factor:

the control factor:

the response variable:

Practice: Researchers are studying the use of a new medication (a tablet taken by mouth once a day) that is designed to lessen the severity of migraines for people who suffer from migraines. The recruiters gather 200 participants to determine whether the medication is effective. Identify whether each is describing a **pre-post design, a randomized controlled experiment, or a non-randomized controlled experiment.** For the multi-group designs, identify whether **blocking, crossover trials, or repeated measures** were used.

Choose the 100 people with the closest addresses to the clinic to be the "treatment" group, and have the other 100 be the control group.

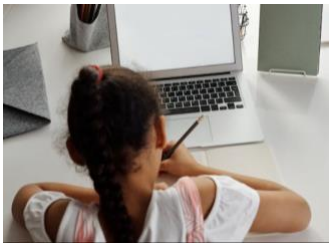
List all names on cards, shuffle them up, and then let the first 100 cards chosen be the treatment group. The other 100 will be the control group. After an initial completing a one month cycle, the groups switched interventions for another month.

Consult demographic information about each participant (sex, age group, race) and randomly assign each subgroup to ensure proportional representation in each experimental condition. The researchers measured participants current migraine levels before and after completing their intervention cycle.

Measure all participants' migraines at beginning of the study. Then have everyone take the treatment tablet for 2 weeks. After 2 weeks, measure migraine levels again.

Practice: An educational researcher is curious whether students in an online course learn more using "directed learning" videos or "active learning" videos. This researcher creates both sets of videos.

Of 139 students who enroll, 70 are randomly assigned to the active learning videos and 69 to the directed learning videos. By the end of the semester, there are 53 students in the active learning group who complete the final exam and 64 students in the directed learning group who do so. She gives both classes the same exam at the end to see which class has improved the most.



The active learning group has an average of 87.8 compared to the directed learning with 86.4. The p-value in this comparison comes to 0.004.

Does this provide evidence that the active learning videos improved performance? Any causality threats?