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. https://mediaspace.illinois.edu/media/t/1_3ae3iepg

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

Population: All (adult U.S.) humans

Explanatory Variable: whether perpetrator was attractive or not

Treatment factor: attractive mugshots

Control factor: regular (non-attractive) mugshots

Response variable: suggested prison time

This study was likely **blinded**. That means participants were…not told which group they were in or what intervention the other group would have.

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?

No, since the same person led both group interventions.

Some studies may use a **Placebo—**a non-effective substance/intervention that are designed to mimic the interventional experience of the treatment factor. In this study, no placebo was necessary since there was an appropriate comparative intervention to directly pair against the treatment factor.

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.

Pre-Post Designs: All units complete/undergo complete the same task(s)/intervention(s) in the same order. We then compare each set of measurements 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?

- Placebo Effect: Psychologically feeling more relaxed because they know they are receiving something. We can't separate the tablet's substance from the experience of receiving a sleeping pill.

- Reactance: Knowing that they are taking the sleep tablet, participants may make other changes consciously or unconsciously that affect sleep (diet changes, bedtime changes)

- Timing Effects: If all taking sleep aid same week, it's possible that timing-related events may be systematically different between those two weeks. Weather differences, current events, etc.

- **Investigation (Reward vs. Punishment):** Another *100 Humans* experiment examined under what conditions humans perform better: https://mediaspace.illinois.edu/media/t/1_z9k92qkb

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

Test Familiarity: Participants might have performed better the second time since they had a practice run

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

Placebo Effects, Reactance, Timing Differences, Test Familiarity

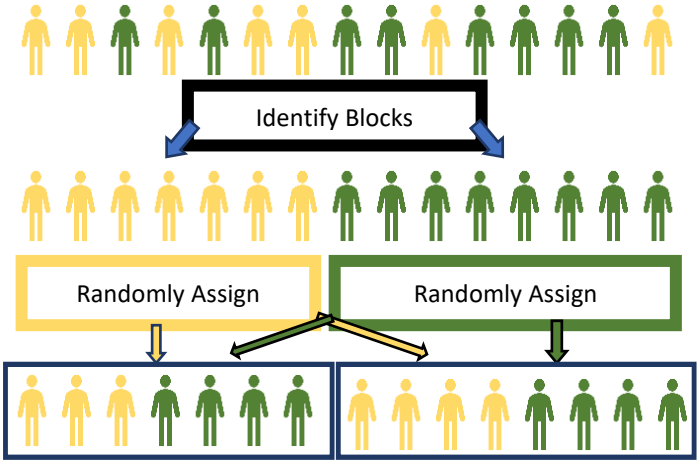
Multi-group designs: In a multi-group design, we can now separate the treatment/control factors into separate groups and potentially avoid other confounding differences, such as timing differences, test familiarity, or reactance/placebo effects. There are *several* design types and features we'll discuss.

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 <u>equivalent</u> 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 <u>confounding</u> 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 <u>small.</u>



- **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
<u>Non-random (confound by race/eth?)</u>	<u>Random (no blocking)</u>	<u>Random with blocking</u>

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

1. In a pre-post design, we record the before and after data for each participant. This can give us more precise data on true average differences and more data on individual variation.

- **Randomized Controlled Experiment with *Repeated measures***
    - If both experimental conditions involve some type of intervention, then we might choose to gather that additional data from each unit!
    - Note that it may not always be advantageous to take pre-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 pre-post design, we can get treatment factor responses from all of 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 still maintain two (or more) independent groups, but have each group complete 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. In some studies, researchers may add some buffer time between each phase.

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

Randomized Controlled Experiment: Half do reward and half do punishment. Compare.

…with Crossover trials. Half start with "reward" and half start with "punishment," and then switch and have them do the other.

Investigation (Mathbar): A randomized controlled experiment assessed if students' mathematical performance was enhanced by taking *MathBar*. 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 at 10am. 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 evidence that MathBar causes an increase in mathematical performance, or are there some possible confounders present?

- Different instructions: mathbar told it would boost performance—perhaps psychological difference
- Mathbar vs. eating/receiving something in general (have we isolated the right thing?)
- Room differences (setup, temperature, different proctor)

- **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?
    - **Attrition** is a threat when participants drop out in different rates between our groups, or drop out for different reasons. *Those who stay may be biased toward the stronger/more capable ones, as they are the ones who stuck through.*
    - Participants may even pass away (**Mortality**), perhaps even as a result of the condition being treated or the treatment itself.
    - Drop out differences become a threat whenever 1) drop out rates are different between groups or 2) drop out reasons are different between each group.

  - **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 itself!

    - This problem is exacerbated in a Pre–Post Design when there is no control group to compare that test familiarity bump to.
    - Test familiarity typically cancels out in multi-group designs, but big test familiarity bumps may mask the true effect of the treatment!

- **Timing Effects** – Do systematic differences in group timing affect outcomes?
  - Are the comparative response measures taking place at different times/days?
  - 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)

- **Setting Effects** – Do any other setting or experiential differences affect the response?
  - **Reactance:** People perform differently because they know they are being studied. *This is especially a threat in pre-post designs.*

  - **Placebo Effect*:*** Are participants improving just because they know they are receiving something. *This is a concern when we **don't** 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 Double-Blinding 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? Staff differences? Room differences? Location 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 1.

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 take a look at one more clip from *100 Humans.*
https://mediaspace.illinois.edu/playlist/dedicated/1_tw9nkdr5/0_grf97xge

Definitely an independence violation as humans might have been influenced by statements or raised hands of the rest of the group (reverse gender bias masking actual gender bias??)

Researcher effects (same researcher—also asked one of the questions differently with different tone)

In general—this research question is soooo broad. These two Jesse's can't possibly embody all women/men, and how they are individually perceived greatly affects this comparison. Maybe "girl" Jessi has better ratings because she appears more "boyish" than typical women. Male Jesse is very ordinary guy in comparison.

Chapter 10 Additional Practice

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 experimental units: The 100 athletes

the treatment factor: Gatorade

the control factor: Water (this is not a placebo, but it is being used as a comparative factor)

the response variable: Amount of time athletes hold the cycling pace

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.

Non-randomized controlled experiment

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.

Randomized controlled experiment (with crossover trials)

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.

Randomized controlled experiment (with blocking, with repeated measures)

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.

Pre-Post Design

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?

p-value suggests difference is unlikely due to random chance. But…

Drop-out differences are suspicious, as the active learning group had a much higher drop-out rate (was that section harder? Are the ones who stayed biased toward better students?)

Possible "researcher effect" in that researcher may have preferred one or built the exam to better match one.

Possible Independence violation if members of the section interacted. Group dynamics at play?

Chapter 10 Learning Goals

After this chapter, you should be able to…

- Recognize the strengths of experimental design to identify causality through its use of controlled interventions
- Define and experimental terms and identify them in an experimental description
    - Treatment factor, control factor, placebo, blinding, double blinding
- Identify the treatment factor(s), control factor (if applicable), and response variable in a described experiment
- Recognize the causality threats common to pre-post designs
- Recognize the value of random assignment (as compared to non-random sorting methods)
- Understand the use of blocking before random assignment and how it is useful with smaller groups
- Distinguish between pre-post designs, randomized controlled experiments, and non-randomized controlled experiments
- Recognize multi-group designs that use repeated measures or crossover trials, and why such features are sometimes useful
- Evaluate a study's causality argument through the lenses of group selection, setting, timing, drop-out differences, test familiarity, and independence
- Identify design changes that improve a study's statistical power
- Evaluate a design by identify weaknesses to the causality argument and recognize design changes that might improve the causality argument