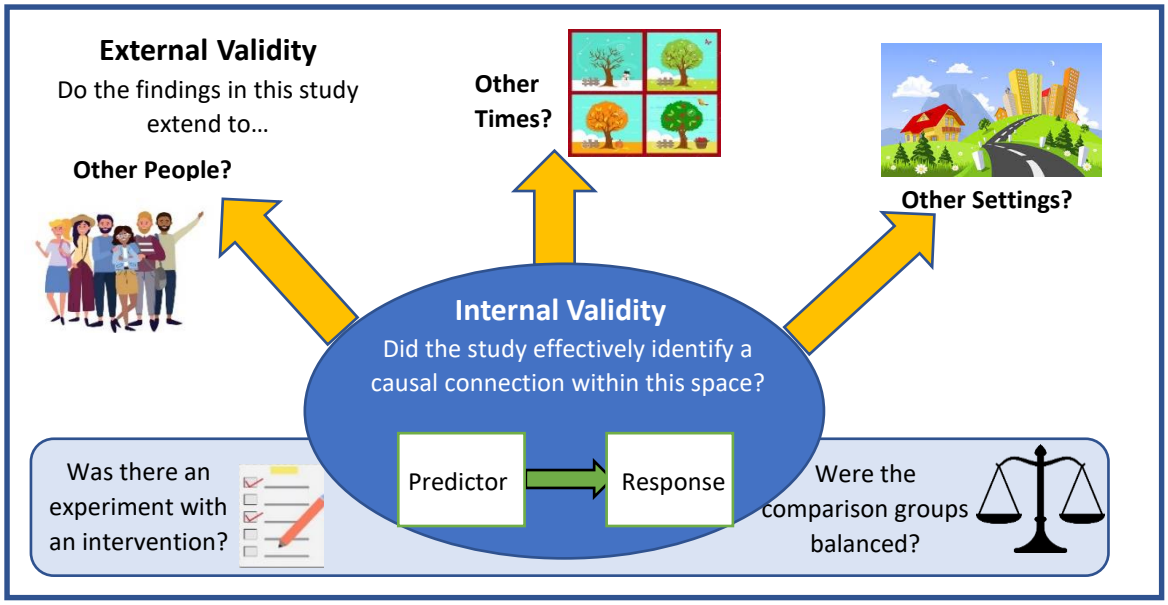
Chapter 10: Evaluating Internal Validity

Identifying Internal Validity

- As a reminder, validity can be divided into two categories:
  - **Internal validity**: Is there evidence for a **causal** link between two variables within this study? Think looking "internally."
    - Does smoking *cause* cancer?
    - Does this medication *directly decrease* LDL cholesterol levels?
    - Internal validity is relevant to assess when exploring *multivariate* questions.
  - **External validity**: Is there evidence that findings in this study **generalize** to a broader population, setting, and time? Think looking "externally."
    - We surveyed 500 people, and 56% approve of the President's performance. How well do these 500 people *represent* the greater U.S. population?
    - External validity is relevant to assess for both univariate and multivariate questions.
  - The **design** of a statistical study helps us determine a study's *internal validity*.
  - The **design** of a statistical study helps us determine a study's *internal validity*.
  - The **sampling procedures**, **setting**, and **timing** help us determine a study's *external validity*.

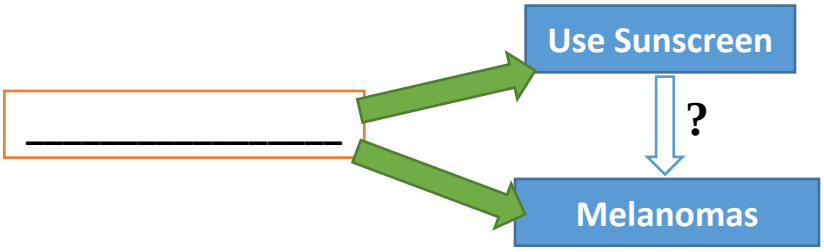


Internal Validity – Evaluating Causality

- Introducing Causality
  - Stratifying is an analytical technique that helps us decide if there may be an association between two variables.
  - But…we need context of how the data was collected or additional analysis before we can decide if changes in one **cause** changes in the other.
- Confounders and Mediators
  - When finding an association between two variables, there are a few possibilities.
  - It's also possible that *more than one* of these could be true at the same time!

Causality	**No Causality**
• The predictor directly causes changes in the response • The predictor indirectly affects the response variable by beginning a causal chain that will affect the response (there is a mediating variable)	• The predictor is merely associated with something that causes changes in the response (there is likely a confounding variable)

Example of Confounding Variable. Consider a medical study to examine factors that might lead to melanomas (skin cancer). One researcher notes that people with melanomas were much more likely to have reported using sunscreen in the last year. Does that mean that sunscreen is causing skin cancer? Can you think of any possible confounders in this relationship?



People with lighter skin or people who go out in the sun more probably wear more sunscreen. The sunscreen is likely not causing the melanomas—it's just associated with factors that do!

Example of Mediating Variable: People who earn more income tend to have longer lives. Does that mean that money itself is directly increasing lifespan?

More likely, wealth begins a causal chain that is mediated by better healthcare or better quality of living. Perhaps also worth noting that we don't often care too much in distinguishing direct vs. indirect causality. Just focus on distinguishing causality from confounding.

Design as a Critical Consideration

- Before assessing a study's internal validity, we should first identify what type of design was used.

Experiments	**Observational Studies**	**Descriptive Studies**
• Designed to identify *causal* relationships. • Experiments involve the assignment of experimental units to an intervention.	• Identify *associations* that may or may not signal causation. • Observational Studies don't assign participants to interventions, but instead observe different variables and identify relationships.	• *Not* designed to identify relationships between variables. They focus on answering univariate questions. • Descriptive Studies typically report summary statistics and inferential results for individual variables

- **Experiments**
  - **Treatment factor:** A factor being investigated as a possible cause. It is the intervening agent.
  - **Control factor:** A comparison factor that is used as a baseline to understand the treatment factor. May be a placebo, a standard treatment, or some well-understood factor.
  - **Response:** A variable of interest that may be the "effect" from the "cause" identified above.

    Examples: A new medication → Higher recovery rate?
    Examples: Smoking → Lung cancer?

  - **Treatment group:** The observational units assigned to the treatment factor
    Note: an experiment may have more than one treatment group!
  - **Control group:** A comparison group that receives a control factor (or no intervention).
  - **Placebos** are typically used to mimic the psychological effect of the treatment without the supposed, unseen benefits of the treatment.
  - **Blinding** means that participants do not know if they are in a treatment group or control group, and **Double Blinding** means that staff interacting with participants also doesn't know!
  - **Random Assignment** means that participants were sorted randomly to experimental groups.

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

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!

First Consideration: Did the Experiment use Multiple Groups?

Developers of a new sleep aid study its effect on improving average duration of sleep. Does one design leave the door open for confounders more than the other?

One Group with Pre-Post: Participants report how much sleep they get on average for the week before taking the sleep aid. They then do the same thing for a week while taking the sleep aid. Researchers compare data from the week before the sleep aid to the week of using the sleep aid to determine if weekly sleep results went up or not while on the sleep aid.	**Two Groups with Post only:** Half the participants are randomly assigned to take a placebo tablet, while the other half are randomly assigned to take the real sleep aid tablet. Each group reports how much sleep they get for the week they take their assigned tablet. Researchers compare data from each group to determine if weekly sleep results were higher or not for the sleep aid group as compared to the placebo tablet group.

The one-group option creates lots of possible confounders

- 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.

- Psychologically feeling more relaxed because they know they are receiving something (placebo effect). We can't separate the tablet's substance from the experience of receiving a sleeping pill.

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

If having pre-measures sounds advantageous, it's because it is! Just not with internal validity. It will come up again later.

Second Consideration: Were groups comprised equivalently?

Do you think one sorting method might be more effective than the other in creating equivalent groups?

Sorting Method A: The first 50 participants to sign up for the study were assigned the treatment group, while the second 50 were assigned the control group.	**Sorting Method B:** Researchers used a spreadsheet to choose participants from a list at random to place into the treatment group. After 50 had been chosen at random, the other 50 were assigned to take the real sleep aid tablet.

Sorting Method A may have a systematic bias in its group composition. People who sign up first may be more motivated or desperate. This may result in non-equivalent groups

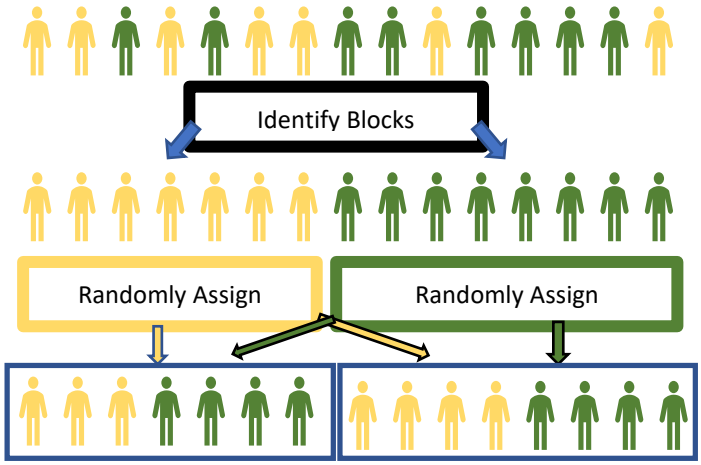
Sorting Method B leaves differences up to random chance

- Experimental Design Options
  - **Pre-Post Design**: All participants have response measurements taken before and after the treatment intervention.

Pre-measure → Treatment → Post-measure

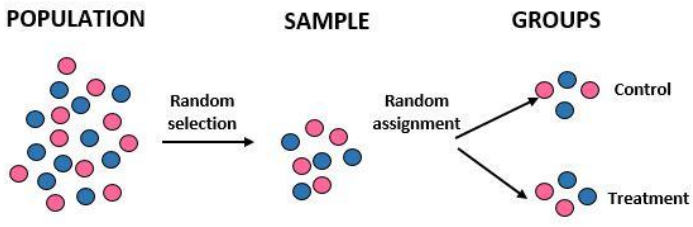
    - Pre-post designs have *poor internal validity* due to lacking a comparison group.

  - **Randomized Controlled Experiment:** Uses pure random assignment or random assignment with blocking to sort groups
    - ***Random Assignment*** means that the researchers sort participants using a random mechanism *(coin flips, computer randomization, names from a bowl, etc.).*
    - ***Random Assignment with Blocking*** does not use *pure* random assignment, but first involves blocking participants and splitting them up to each group equally.
      - ❖ 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 participants in each block to a group.
      - ❖ Blocking is generally good practice when groups are *small* (e.g., sizes are < 50).



    - If the researchers wish to take pre and post measures to measure response variable changes, this would be a Randomized Controlled Experiment ***with Repeated Measures***.

    - ***Random selection*** *and* ***random assignment*** *are* ***not*** *the same thing.*
      - ❖ Random assignment refers to the sorting of experimental units into treatment and control groups once observational units for the study have *already* been selected.
      - ❖ Random sampling involves random selection of observational units into the study to begin with.



Statology (2020). https://www.statology.org/random-selection-vs-random-assignment/

- **Non-randomized Controlled Experiment**: Uses a non-random assignment method to sort groups
  - ***Non-random assignment*** could create a systematic bias in groups and should be used or assessed with caution. Examples:
    - ❖ Arranging people by last name alphabetically. First half go to treatment group and second half to control group
    - ❖ Sorting people by their birth month—January-June to one group, July-December to the other group.
    - ❖ How might these sorting methods create bias?

Last names: ethnic imbalance.
Birth month: school starting age, or current age imbalance

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. **What type of design is used in each situation?**

Choose the 100 people with the closest addresses to the clinic 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.

Randomized controlled experiment (pure random assignment)

Consult demographic information about each participant (sex, age group, race) and randomly assign each subgroup to ensure proportional representation in each experimental condition.

Randomized controlled experiment (with blocking, then random assignment)

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

Third Consideration: Were group experiences similar in all ways except for the treatment factor?

A randomized controlled experiment was conducted to assess if students' mathematical performance was enhanced by taking *MathBar*, a new 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 at 10am. The other half did not receive anything and served as a control group. They completed their exam in another room at 3pm. 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.
- Room differences (setup, temperature, different proctor)
- Different times

- **"Internal Validity Threats" to consider for Experiments**
  - **GROUP SELECTION** – Have the groups been sorted equivalently?
    - 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 is a possible threat whenever the response variable is measured both before and after treatment.
    - This threat would be most pertinent when the instrumentation is a duplicated mental or physical test. Participants' results may change due to familiarity and improvement rather than treatment. *They 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.

- **TIMING EFFECTS** – Do systematic differences in group timing affect outcomes?
  - Are the groups' interventions taking place at different times? On different 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** – Is the setting and experience equivalent for all groups?
  - **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 (more encouraging or more engaged in the interaction, etc.). 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?

Practice: An educational researcher is wondering whether her students learn better in a traditional classroom setup or a flipped classroom setup (Flipped classroom is when students watch videos as homework and come to class for activities and problems). Section A is a traditional section with 74 students and Section B uses a flipped format with 89 students. By the end of the semester, Section A has 67 students, while Section B has 68. She gives both classes the same exam at the end to see which class has improved the most.

What are some internal validity threats?

Group Selection is a threat (unclear if these groups are equivalent. Likely no intentional sorting)

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

Setting could be a threat

- Different environments (classrooms)?
- Researcher effect (did prof treat each class differently?)

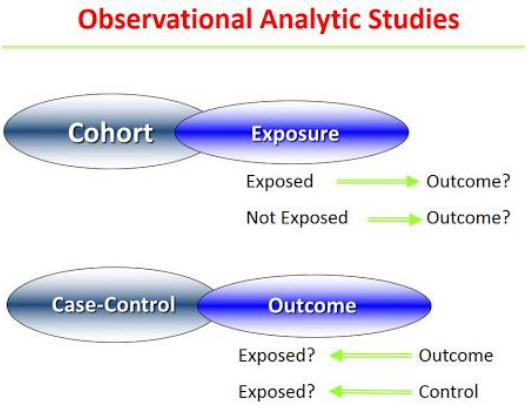
Timing could be relevant if classes were at different times (unclear)

NOT THREATS: Different group sizes does not hurt internal validity. There is no test familiarity.

- **Observational Studies – When Experiments are not Reasonable to do**

  - Observational studies look at the relationship between two variables, but without assigning participants to different interventions.
  - In some cases, it is not possible, or not ethical, to assign participants to different interventions.
    - Studying the potential effects of alcohol consumption on fetal development. Not ethical to assign pregnant women to alcohol
    - Studying whether blood lead levels in children are different depending on their neighborhood. Not possible to assign children to live in a different neighborhood
  - Observational studies simply observe life as it would already proceed. Without controlled intervention, there will always be possible confounders that could explain the correlation.

- Three common observational study formats:
  - **Cross-sectional Studies**
    - Cross-sectional studies collect data during a specific and limited period of time. It's **data at a cross-section** of a participant's life.
    - Cross-sectional studies are commonly in the form of a one-time survey. Each participant only logs their data for one time point.
  - **Cohort Studies**
    - Cohort studies involve identifying several groups (cohorts) of people based on a shared characteristic. Researchers then wait some period of time to see if each group has a difference in outcomes.
    - Cohort studies are commonly Prospective in form, meaning that we are "looking ahead" to measure the response variable in the future based on a naturally-occurring "treatment factor." (e.g., effects from alcohol, smoking, having a certain gene).
  - **Case-Control Studies**
    - In case-control studies, researchers identify people who have had a certain response, and then look to see if there are any common characteristics that might explain that response.
    - Case-Control studies are commonly Retrospective in form, meaning we have identified a response, and are looking back in participants' past to find a potential causal agent.



Identify whether each design below is an observational study or an experiment. If obs study, what type?

A survey is conducted to college students asks whether they eat dinner at approximately the same time every day. This survey also asks how many hours of sleep they get a night. The team is curious if people who eat dinner at a regular time also get more sleep on average.

Observational study (cross-sectional). Dinner eating habits in this study are indicative of their natural habits

In another variation of this investigation, researchers took a group of students who did not normally eat dinner at the same time and randomly chose some of them to choose a regular dinner time for 2 weeks. The others continued with life as normal. At the end of 2 weeks, the researchers compared the sleep amounts of those who stuck with the regular dinner time to those who continued without any change.

Experiment. Students assigned to each option

To determine how effective masks are in preventing the spread of COVID-19, researchers identified cities that had a mask mandate and cities that did not. They then tracked the percentage of residents who contracted COVID-19 over the following 4-month period.

Observational study (cohort). Tracking the results of city decisions—there could be other reasons that explain differences in infection besides mask mandates.

In a lab setting, researchers had participants who had a recent diagnosis of COVID-19 do various things. Some were provided a mask, while some were asked not to wear a mask. A device measured levels of the virus suspended in the air after each activity to see whether masks affected viral load in the air.

Experiment. Participants assigned to wear a mask or not.

A group of cardiologists identified patients with diagnosed heart disease. The researchers then looked back at medical records to determine which were prescribed a particular aspirin that the researchers suspected might have links to heart disease.

Observational study (case-control). The response variable is measured first. We then identify a possible causal factor.

Improving Validity versus Improving Power

- As a reminder, **Power** is concerned with how much power our study has to detect a departure from the null hypothesis if the null is false. We can improve power in several ways:
    - Increasing the sample size. *(increase sample size, reduce standard error!)*
    - Decreasing "random noise" in measurements. *(reduce variation, reduce standard error!).*
        - Think more precise instrumentation, more standardized procedures, or by using repeated measure designs. Make the signal easier to see through the noise!
    - Having equal group sizes can *slightly* improve power. If I have 100 people, I'll have more statistical power comparing groups of 50 and 50 than I would with 80 and 20.

Practice: Does drinking caffeinated coffee make students more productive? To test this, a researcher gathers 60 college students who don't normally drink coffee. 30 of them are assigned to a decaf coffee in the morning, while 30 are assigned to drink caffeinated coffee in the morning. At the end of the day, students provide a self-report on their productivity from 1 to 10.

Consider the following design options and consider whether it affects the study's internal validity, external validity, power, or multiple of these!

Since there are only 30 students in each group, the researcher decides to block by other relevant factor, like gender, other caffeine habits, and workload.

Improves internal validity (likely more equivalent groups)

The researcher decides to turn this into a large-scale survey with several hundred college students. Students are now asked to report their coffee-drinking habits for the previous day, as well as self-report their productivity for the previous day.

Improves power, but definitely hurts internal validity. Introduces lots of confounders!

The study was originally going to have all students buy the caf or decaf coffee of their choice, as well as choose the amount to drink. Instead, the researcher decides to give all students the same brand and prescribe the same amount.

This is a hard one! Mostly, this is improving power by standardizing the instructions. Too much volatility in choices may make the signal hard to see! Mildly hurts the external validity, as this is now specific to one brand and one dosage, rather than just coffee in general. But this is probably a good tradeoff! It's ok to make specific generalizations!

Chapter 10 Learning Goals

After this chapter, you should be able to…

- Distinguish questions of internal and external validity
- Distinguish between confounders and mediators
    - Identify a confounder as a possible causal explanation to the association between a predictor and response.
    - Identify a mediator as a possible causal link between a predictor and a response
- Distinguish between experiments, observational studies, and descriptive (univariate) studies
- Recognize internal validity as generally stronger in experiments, generally weaker in observational studies, and not applicable to descriptive studies
- Define and experimental terms and identify them in an experimental description
    - Treatment factor/group, control factor/group, placebo, blinding, response, random assignment
- Recognize how internal validity may be weakened in experiments that use only a pre-post design with no control group
- Recognize how internal validity may be weakened when a random assignment is not used to sort participants to experimental groups
- Distinguish between pre-post designs, randomized controlled experiments, and non-randomized controlled experiments
- Define “blocking” and “repeated measures” in the context of a randomized controlled experiment
- Evaluate a study’s internal validity through the lenses of group selection, setting, timing, drop-out differences, and test familiarity
- Recognize situations in which an observational study may be more appropriate—or simply the only option—in certain contexts
- Distinguish between cross-sectional designs, cohort designs, and case-control designs based on an observational study description
- Recognize the tradeoff between internal validity, external validity, and power in designing studies—there are no perfect studies!