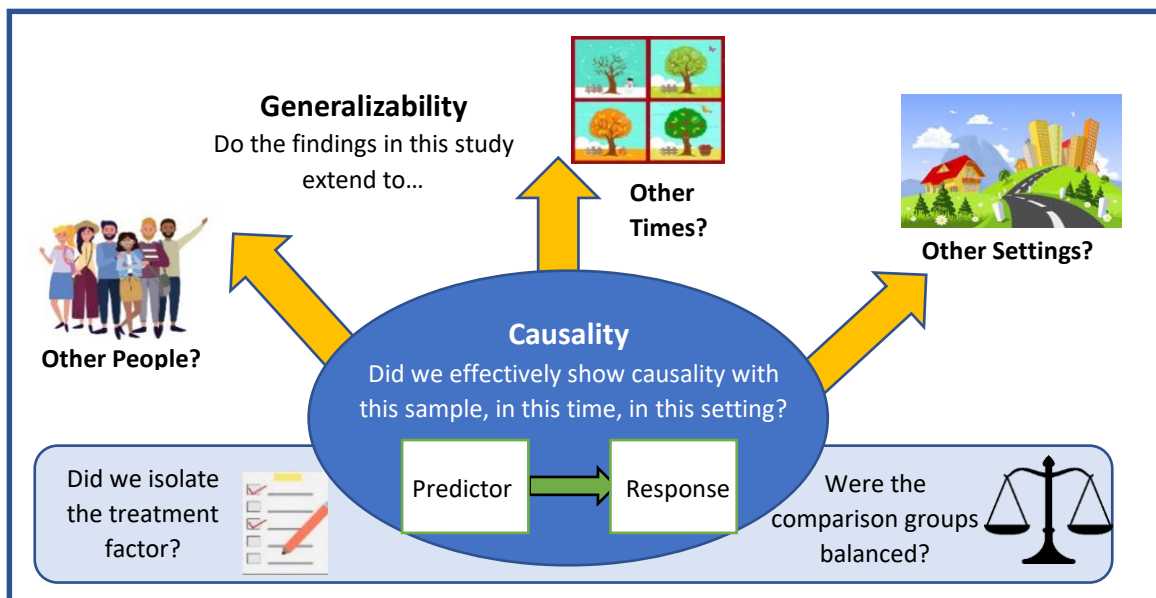


## Chapter 11: Evaluating Generalizability

### Generalizability vs. Causality

- A review of **causality** arguments
  - As we discussed in the previous chapter, causality arguments ask whether a factor from an explanatory variable affects change in a response variable.
    - Does taking this medication *cause an increase* in serotonin levels?
    - Does this vaccine *cause a decrease* in the likelihood for people to contract this virus?
  - We can typically draw stronger causality arguments from \_\_\_\_\_ than we can from observational studies. Do you remember why?
- Introducing **generalizability** arguments
  - Generalizability arguments ask whether findings in our study \_\_\_\_\_ to a broader population, setting, and time. Think looking “externally.”
    - Is this medication effective at increasing serotonin levels for adults of all ages?
    - Is this vaccination effective at decreasing viral contraction with new variants of this virus?
    - Did the group of people we surveyed for this presidential approval poll reflect the broader U.S. population?



**Practice:** Which of these questions is targeting the causality argument? Which is targeting generalizability?

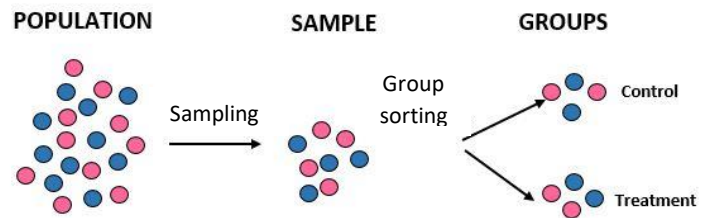
Is taking notes actually leading to more learning? Or is it just that students who take notes happen to be better learners in other ways?

If we repeated this study in a different section of the course with a different instructor, would we see the same result?

Evaluating a study's **sampling process** is a key part to evaluating a study's generalizability. What have the researchers done to ensure their sample represents the population they wish to generalize to?

**Sampling and group sorting are not the same!**

- Sampling is the process of choosing units from the \_\_\_\_\_ to comprise the sample.
- In experiments, we will take one further step of assigning units from the \_\_\_\_\_ to be in one of several groups.



[Statology \(2020\).](#)

A **Simple Random Sample** (sometimes just called

“*random sampling*”) is the ideal method of sampling in large-sample studies. Two things should be true...

- Every member of the population has an **equal** chance to be chosen.
- Sampling remains **independent**: the possibility of one person being chosen does not affect the chances that someone else is chosen.

Unfortunately, random sampling may not be possible due to sampling biases that are difficult to avoid.

- \_\_\_\_\_ **Bias**: Some in the population don't have an equal chance (or no chance at all) to be selected or contacted.
- \_\_\_\_\_ **Bias** (*may also be further distinguished into Self-select or Non-response Bias*): Once being selected, those who *chose* to participate may be systematically different than those who chose not to participate or (forgot/unable to respond).
- In bio-medical studies, be on the lookout for the role of **survivorship**. That is when our sampling process results in an inordinate number of “successful” cases that doesn't represent the population (or even the opposite—an inordinate number of unsuccessful cases!). This can accentuate an undercoverage bias, volunteer bias, or both!



**Practice:** A doctor is examining which forms of anesthesia seem most effective at reducing pain and easing recovery in the days after open heart surgery. He collects data from the procedures completed at a nearby hospital and talks to the patients to gather data about their pain levels and experiences.

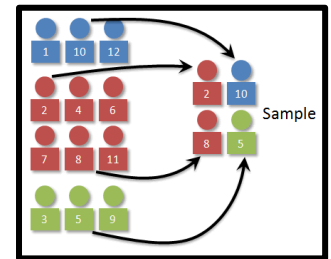
**How might this design create an undercoverage bias**

**How might this design create a volunteer bias?**

**Does survivorship play a role in accentuating any sampling biases?**

Outside of special situations where researchers can take a simple random sample (or some other “probability-based” samples), we can generally describe most every human-subject study as having a **convenience sample**.

- **Convenience sampling** simply means that we have a \_\_\_\_\_ sampling plan.
  - The researchers have gathered a group of participants that was convenient to contact or from whomever agreed to participate and complete the study.
  - There are two common strategies that researchers use (especially public opinion pollsters) to create a better argument for representation.
    - **Quota sampling** identifies key features that we wish to have proper representation of (e.g., age groups) and attempts to sample a target number from each subgroup.
    - For example, if 30% of the adult population is between 18-35, 40% is between 36 – 55, and 30% is 56 and above, then I will attempt to replicate those percentages in my sample.
    - **Weighted sampling** attempts to use a mathematical model to weight responses if that person comes from a subgroup that is over or under-represented in my sample.
    - For example, if 30% of the adult population is between 18-35, but only 15% of my sample is between that age, then I might double the weight of those respondents’ answers by a factor of 2 when calculating summary statistics.
  - But even without these sampling techniques, we can still ask to what extent a study’s sample represents key features of the broader population.



**Practice:** A tech company has developed a smartphone app that uses artificial intelligence (AI) to detect melanoma and other skin cancers from photos of moles or skin lesions. The company shares results from over 10,000 people revealing the app to be highly accurate at detecting these issues. Which subgroups might be important to check for appropriate sample representation before we should generalize these results?

#### Bottom Line about Sampling

- Just as we should be cautious about making causality arguments from observational studies, we should also be cautious to make population generalizations from a \_\_\_\_\_.
- But we can still build an argument for generalizability if we can show that our sample \_\_\_\_\_ the population across demographics critical to the study’s context.

## Threats to Generalizability Summarized

- **Participant Selection** – Does this group of participants represent the population?
  - How might our sampling plan undercover certain key groups? Who is more likely to volunteer? Who was more likely to drop out early?
  - Overall, what key demographics should we check representation of in our sample?
- **Setting Limitations** – Is the setting representative of all settings we wish to generalize to?
  - \_\_\_\_\_ **Environment:** Might the physical space that this study took place in affect any outcomes we found?
  - \_\_\_\_\_ **Environment:** How might any social interactions or specific people involved have affected the outcomes?
  - \_\_\_\_\_ **Features:** What other contextual factors did this study take place within? A particular weather event or season? Specific materials? Dosage? Instrumentation?
- **Historical Sustainability** – Do these results generalize to other times?
  - This threat should be considered when dealing with external factors that may change over time—questions linked to culture, lifestyle habits, entertainment, etc.
  - For example: A poll about Americans’ views about government surveillance or terrorist prevention before the terrorist attacks on September 11<sup>th</sup> 2001 may no longer generalize to Americans’ views after that event.

**Practice:** Researchers in the 1980s were examining the relationship between Americans’ political views and whether they watched news programming on television. The research team contacted residents in New York, Chicago, and Los Angeles, asking to speak to “heads of household”. The researchers concluded that people who more regularly watched the news were more likely to have moderate political views as opposed to non-regular news watchers.



What might be some threats to this study’s generalizability?

**Participant Selection Threat?**

**Setting Threat?**

**Historical Sustainability Threat?**


## Causality, Generalizability, and Power

- In Chapter 1, we talked about three general types of uncertainty in statistical investigations.
  - Uncertainty in **Causality**: This concerns whether we have \_\_\_\_\_ a causal effect. What other systematic differences exist between the groups we're comparing?
  - Uncertainty in **Generalizability**: This concerns \_\_\_\_\_. Who's not here? What other aspects of our results may not generalize to life more broadly?
  - Uncertainty in **Precision**: This concerns \_\_\_\_\_. How much error should we expect in our statistic as an estimate for the parameter?
- The term **statistical power** refers to how much we have reduced our uncertainty in precision—Do we have enough “power” to detect an effect if there is one?
- Mathematically, we improve our study's power by \_\_\_\_\_ the standard error of our sample statistic. For example, in a two-mean comparison, that would be:

$$SE_{(\bar{x}_1 - \bar{x}_2)} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Looking at this formula, there are two key ways we can improve power:
  - Increasing the \_\_\_\_\_.
  - Decreasing random sources of \_\_\_\_\_ in our measurements
    - For example, using precise instrumentation, more standardized procedures, or by taking repeated measures.
  - One additional factor that has a *small* effect on power is dividing our sample size \_\_\_\_\_ to each group in an experiment. (e.g., small group size differences are negligible though).
- Some design changes *could* affect a study's causality argument, generalizability argument, and power—design is often a trade-off between these factors!

**Practice:** Imagine if we wanted to know whether a dose of caffeine truly makes students more productive. Our first idea is to find 200 college students and mark down what sources of caffeine they have today. Then at the end of the day, self-rate their productivity on a scale of 1 to 10.



What changes could we potentially make to this study to improve the causality argument, the generalizability argument, or its power to detect a correlation?

Causality Argument Improvements	Generalizability Argument Improvements	Power Improvements
---------------------------------	--	--------------------

### Reflection Questions

---

**11.1.** How is a threat to a study's generalizability argument different from a threat to its causality argument?

**11.2.** When conducting an experiment, what is the difference between sampling and group sorting? Which more clearly affects a study's causality argument? Which more clearly affects a study's generalizability?

**11.3.** What would have to be true for sampling to be "simple random sampling"?

**11.4.** Are simple random samples common in studies involving human populations? Why or why not?

**11.5.** What are the two common types of sampling biases we learned about? What distinguishes them?

**11.6.** What is a convenience sample? What are two strategies we learned that researchers may use to build a generalizability argument in these situations?

**11.7.** In addition to a participant selection threat, what are two additional generalizability threats we learned? What is an example of each?

**11.8.** What does it mean when a study has more statistical power? What are two ways to improve power?

---

**Chapter 11 Additional Practice (Videos available in the Ch 11 module on Canvas!)**

---

**Practice:** Consider each sampling plan. How might the sampling biases we talked about threaten the ability to generalize the results more broadly?

To better understand how much time American adults spend outside on a typical day, a poll on msn.com users how much time they have spent outside that day.

A hospital emails a survey out to all 2,874 patients who had a procedure and overnight stay completed in the previous year to ask about their satisfaction with their visit. 309 (11%) of them respond to complete the survey. According to clinic data, 82% of the respondents who completed the survey said they would recommend the hospital to friends or family.

A university selects 100 graduating seniors by randomly selecting their email addresses from among those who have applied for graduation. These 100 students are asked to complete an exit interview to better understand what supports helped them complete their degree. 75 of the 100 contacted students completed the exit interview.

**Practice:** In February 2020, Pew Research did a poll to gauge how much Americans were planning to travel the following summer. In reality, significantly fewer Americans traveled later that summer. Which threat category do you think best explains why their results didn't generalize well?

**Practice:** We're completing a study to estimate the amount of time that University of Illinois students (of all academic levels/programs) spend on school each week. Consider the following sampling plans: What potential issues can you think of for each that may limit the generalizability of the claims we make from each?

Take a sample of students taking STAT 100 during the Fall

Conduct a poll on the UIUC reddit page.

