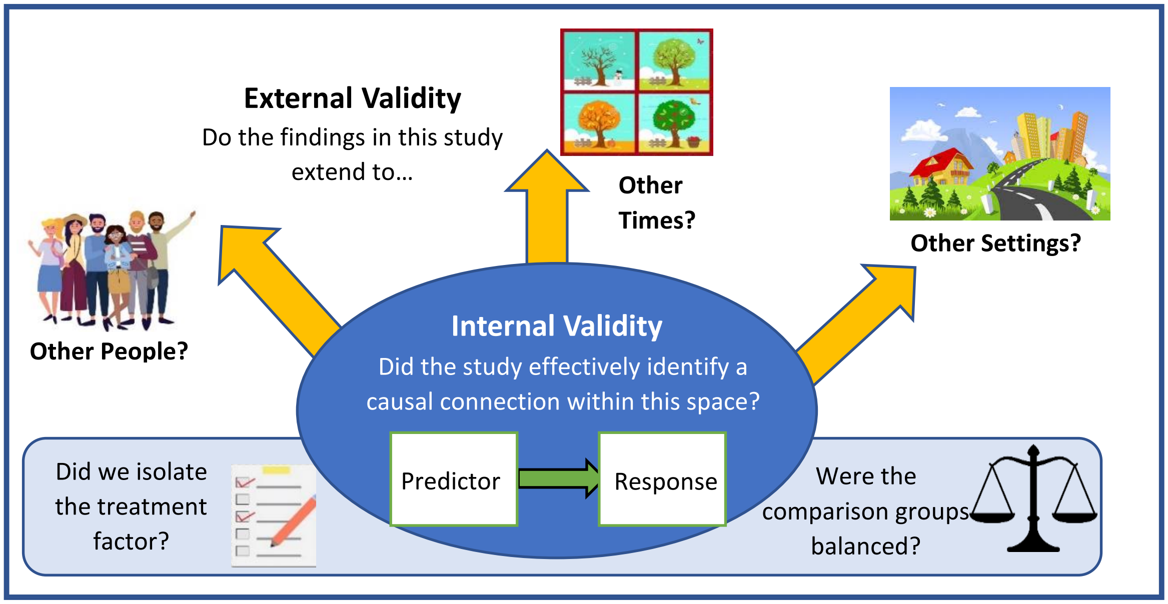
Chapter 11: Evaluating Generalizability

Evaluating Validity

- In the Research Methods world, discussions of evaluating causality and generalizability are referred to as **Validity** arguments.
- 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 **sampling procedures** and overall **setting** and **timing** help us determine *external validity*.


External Validity
Do the findings in this study extend to…
Other Times?
Other Settings?
Other People?
Internal Validity
Did the study effectively identify a causal connection within this space?
Predictor
Response
Did we isolate the treatment factor?
Were the comparison groups balanced?

Practice: One study found that students who took handwritten notes in one particular class performed better on the Final Exam. Decide whether each question below is targeting the study's internal or external validity.

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

External Validity – Determining Generalizability

Sampling is a key part of external validity. In order to determine how generalizable our results are, we need to make the case that our units represent our population of interest.

- A **Simple Random Sample** (sometimes just called "*random selection*") would be the best way to gather a representative sample from the population.
  - Every member of the population has an *equal* chance to be chosen, and sampling remains *independent*: the possibility of one person being chosen does not affect the chances that someone else is chosen.
  - Unfortunately, simple random sampling is often not possible due to inescapable biases in most contexts:
    - Undercoverage **Bias:** Some in the population don't have an equal chance (or more often, no chance at all) of being selected for the sample. That is because it is often difficult to have contact information for every person or unit in the population. It costs time and money.
    - Volunteer **Bias** *(may also be called Self-select or Non-response Bias)*: The sample is composed of subjects who *chose* to participate and others who chose to ignore or forgot. This is expected with human populations since we can't force participation!

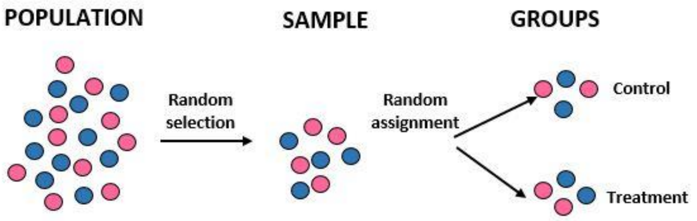
Practice: A pollster is contacting people for a survey on public transportation by collecting responses in a busy downtown square. However, not everyone in the population of interest passes through that square. ***What type of bias would that be?***

Undercoverage (certain people in population have no chance of being asked)

Practice: On the Quad, some students complete the questionnaire, while others decline the questionnaire. ***What type of bias would that be?***

Volunteer (certain people who are invited choose to complete it while others don't)

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

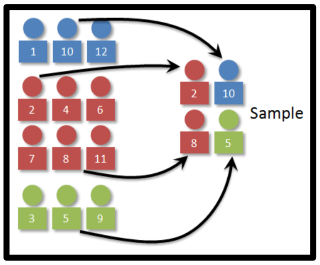


Statology (2020). https://www.statology.org/random-selection-vs-random-assignment/

- **Quota Sampling/Weighted Samples**
    - One way to address undercoverage or volunteer biases is to get target numbers from different key subgroups (e.g., equal male and female, proportional ages, etc.). This is called "Quota" sampling.
    - Examples: Age, Gender, Household Income

    - An alternative to quota sampling is to take a "weighted sample," where we adjust the weight of responses from demographic brackets that are oversampled in order to get what we believe is a more accurate estimate.
    - This type of sampling is often used for public opinion surveys where we know certain subgroups are just harder to contact.

    - Curious to know more? Here is a video on sampling with polls! https://www.youtube.com/watch?v=fzzX9jHDK4k



-
    - **Convenience sampling**
        - This non-random sampling method is quite common in people-centered research. Think online polls or surveys in which there is no attempt to get quotas, or to weight the final responses to represent population demographics.
        - Any time a sampling method has an undercoverage bias or volunteer bias, and that bias is not compensated by some type of weighted or quota sampling scheme, it is technically a convenience sample.
        - People should be cautious of using convenience samples to make statistical generalizations to a larger population.
        - Snowball **Sampling** is a common component of convenience sampling that relies on word of mouth (think shares on social media) to get participants. This would be a *huge* threat to independence in that people participate based on whether they know someone who did. The sample might be an echo chamber.

Threats to Generalizability Summarized

- **Participant Selection** – Does this group of participants represent the population?
    - We need to make the case that our sample **represents** the population at large and identify what elements of our sampling scheme introduce a threat to representativeness.
    - In particular, we should note in what ways undercoverage bias or a volunteer bias may threaten how well our sample represents the population.

- **Setting Limitations** – Is the setting representative of all settings we wish to generalize to?
    - **Physical Environment**: What space or what features were involved in this study. A particular doctor's office? An outdoor vs. indoor location? A biased/limited range of external factors? *Do these features generalize?*
    - **Social Environment**: What is the social context for this study, and might that matter with what was studied: Were particular doctors or nurses involved?
    - **Context**: What other contextual factors did this study take place within? A particular weather event or season? The materials used in the study or instrumentation*?*
    - Note that setting threats to external validity are different than that of internal validity.
        - Setting threats for ***internal validity*** have to do with confounding effects between groups. For example, did my treatment and control group complete their participation in different rooms?
        - Setting threats for ***external validity*** are related to setting encapsulating my whole study. Perhaps my participants completed the treatment in a lab, but would these results generalize to household settings?

- **Historical Sustainability** – Do these results generalize to other times?
    - This threat should be taken into account 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 11th 2001 may no longer generalize to Americans' views after that event.

Practice: Researchers in 1988 were examining the relationship between Americans' political views and whether they watched news programming on television. The research team contacted residents in America's top 10 most populated cities, and callers asked 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 threats to external validity are present in this study?

Participant Selection Threat?	**Setting Threat?**	**Historical Sustainability Threat?**
Residents in large cities may not represent other areas. Men over-represented with "H of H" language	Perhaps local news in big cities is more neutral than in non-urban areas?	The nature of news programming has changed since 1988!

Internal Validity, External Validity, and Power

- Internal and External Validity are more concerned about the quality of our study and the claims we can validly make
  - **Internal Validity** asks whether a relationship we have identified is causal.
  - **External Validity** asks whether the result we found in our data generalizes more broadly
- **Power**, in contrast, is more a question of quantity.
  - Is our study large enough and efficient enough to detect a departure from our null hypothesis? Do we have enough "power" to detect an effect if there is one?
- Mathematically, we improve our study's power by decreasing the standard error of our sample statistic. For example, in a two mean comparison, that would be:

$$SE_{(\bar{x}_1 - \bar{x}_2)} \approx s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

- We can improve power in several ways:
  - Increasing the sample size
  - Decreasing random sources of variation in our measurements
    - For example, using precise instrumentation, more standardized procedures, or by taking repeated measures.
  - If doing an experiment, we'll also gain slightly more power by dividing our sample size equally to each group. (e.g., small group size differences are negligible though).
- Some design changes could affect both power and internal/external validity!
  - Blocking can improve group balance (IV) and improve efficiency (Power)

Practice: Imagine if we wanted to know whether a dose of caffeine truly makes students more productive. Our first idea is to find 100 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 5.

What changes could we make to this study if we wanted to improve the causality argument that caffeine affects your productivity?

Experiment. Randomly Assign students to differing caffeine levels, since students who consume more caffeine may be different in other ways. Isolate the caffeine—maybe a caf vs. non caf drink?

What changes could we make to also improve the study's power to detect a difference confidently?

Larger sample size

More standardized procedure—perhaps focusing on a specific source of caffeine, or having a more consistent way to measure productivity than the 1 to 5 scale. May also improve IV! Also might limit EV a tad.

Chapter 11 Additional Practice

Practice: What do you think of each sampling plan? What limitations might apply to the participant selection in each situation?

> A poll on msn.com asks American users 18 and over whether they plan to vote in the upcoming Midterm elections. After voting, the website encourages people to share the link of the poll with their friends on social media.

Snowball sampling accelerates bias by getting more people with common views (also undercoverage via the audience of msn.com)

> A clinic is surveying the 874 patients from the past year to assess satisfaction with their recent visits. 209 (24%) of them complete the survey. According to clinic data, 68% of the respondents said they would likely visit again or recommend to family members.

High volunteer bias. Those who responded may be more likely to have positive experiences.

> 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 for $20. At the conclusion, a total of 75 of the 100 contacted students completed the exit interview.

Small volunteer bias, but overall likely good representation.

Practice: In February 2020, Pew Research did a poll to gauge how much Americans were planning to travel the following summer. Their results ended up being very off. Which threat category do you think best explains why their results didn't generalize well?

Historical sustainability! Views/expectations in February 2020 didn't generalize after COVID hit.

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 external validity of the claims we make from each?

> Take a sample of students taking STAT 100 during the Fall
>
> STAT 100: certain majors excluded. Possibly biased toward freshmen/sophomores
>
> Conduct a poll on the UIUC reddit page.
>
> Reddit: Biased toward STEM programs, or possibly more introverted types

Chapter 11 Learning Goals

After this chapter, you should be able to…

- Define and distinguish internal validity from external validity
    - Identify “internal validity” as the evaluation of a study’s ability to determine causality between a treatment factor and response.
    - Identify “external validity” as the evaluation of a study’s ability to generalize its findings more broadly
- Recognize and distinguish between an undercoverage bias and a volunteer bias in sampling methods
- Identify a simple random sampling method (aka “random sampling” or “random selection”) from a contextual description and recognize its strong external validity
- Identify quota and weighted sampling methods from a contextual description and recognize their value in minimizing undercoverage and volunteer biases to establish decent external validity
- Identify a convenience sampling method from a contextual description and recognize its relatively weak external validity
- Distinguish sampling (selecting units into a study) from assignment (assigning units to experimental groups)
- Recognize limitations to a statistical claim’s external validity by using the lenses of participant selection, setting limitations, and historical sustainability.
- Recognize questions of a study’s power as relating to a study’s ability to detect a departure from the null hypothesis
- Identify sample size and efficiency features as ways to improve a study’s power