

BEET: Building Evidence Ecosystems about Employability and Skills Development through Training

Training Session 3: How To Randomize and Sample Size

Dr. Reham Rizk, Director, Egypt Impact Lab

Institutional Host



المعهد القومي للتنمية
الاقتصادية والاجتماعية
National Institute for Economic
& Social Development

Founding Partners



SAWIRIS FOUNDATION
الاسرة



Community
Jameel

Additional Support



Course Overview

1. Why Evaluate & Theory of Change?
2. Why Randomize?
- 3. How to Randomize & Sample Size**
4. Generalizability

Lecture Overview

- What is Randomization?
 - Random Sampling
 - Random Assignment
- Randomization Procedures
- The Unit of Randomization
- Sample Size Considerations

Random Sampling and Random Assignment

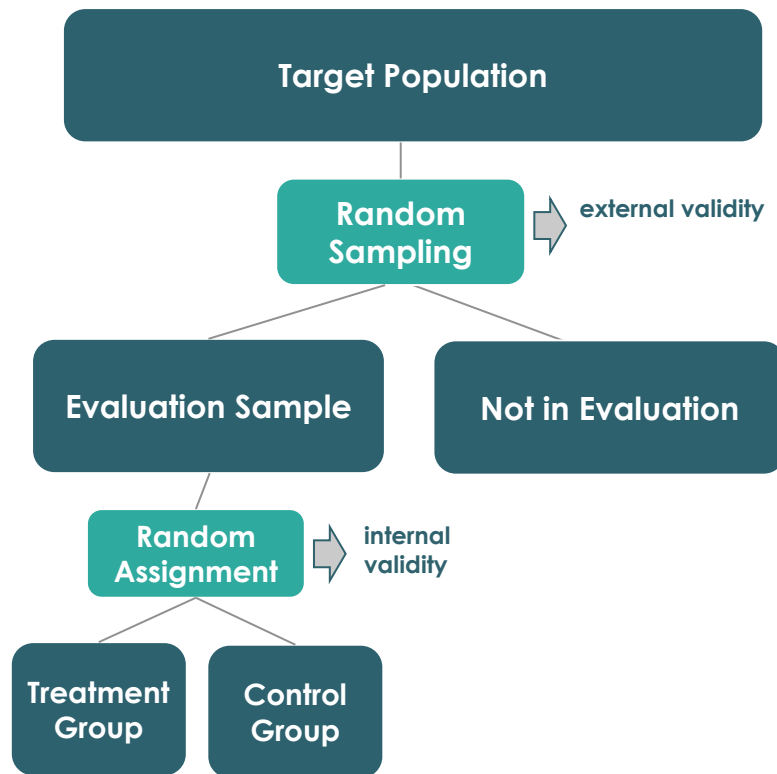
Defining terms

Random sampling: selecting units from a population of interest in a randomized manner to create a sample that is representative of the population

External validity: the acceptability of results of an evaluation in contexts other than those in which the experiment was conducted

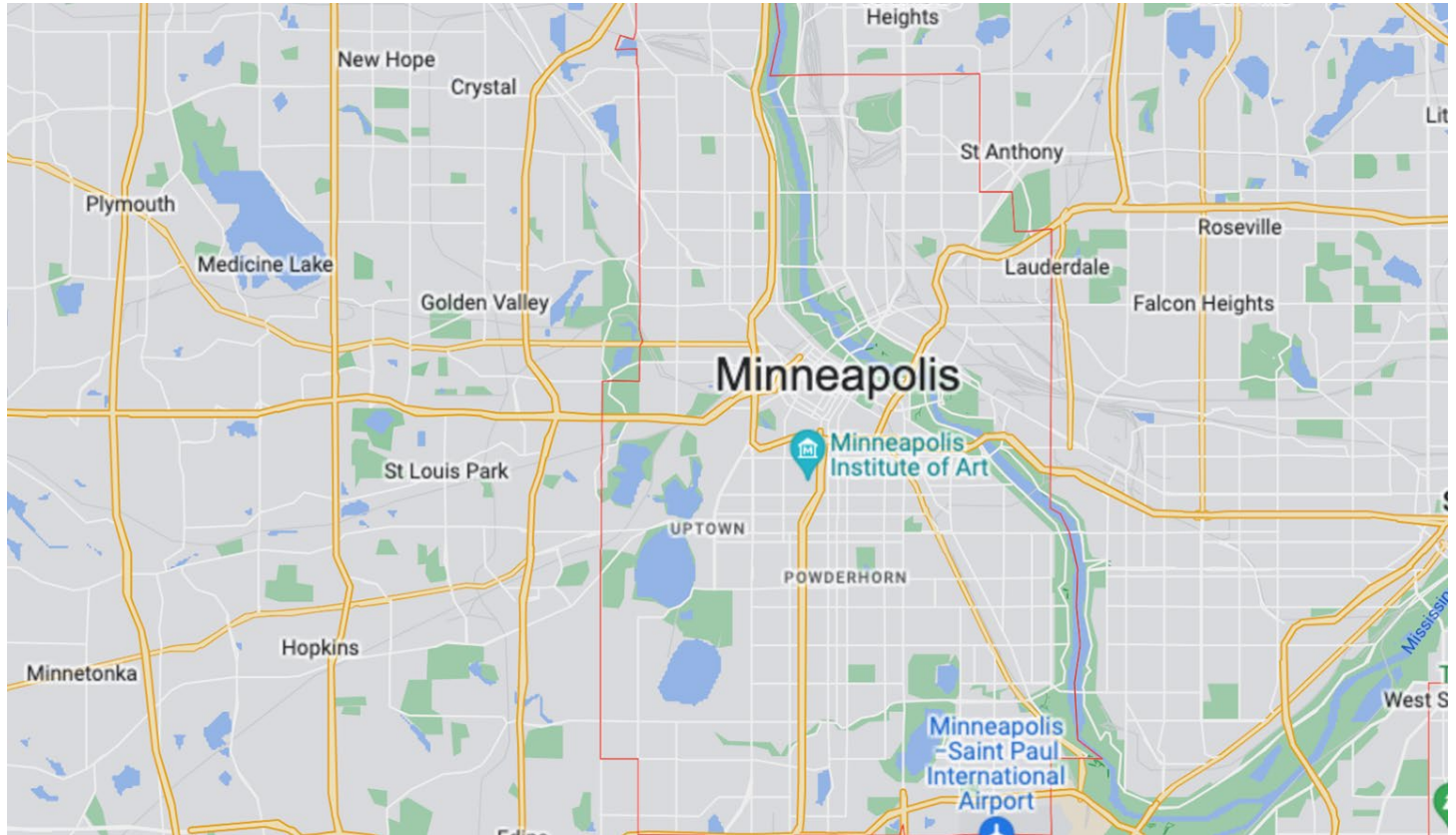
Random assignment: taking a pool of eligible units and then allocating those units to treatment and control groups by means of a random process

Internal validity: the acceptability of the results of an evaluation in terms of causal impact of the intervention



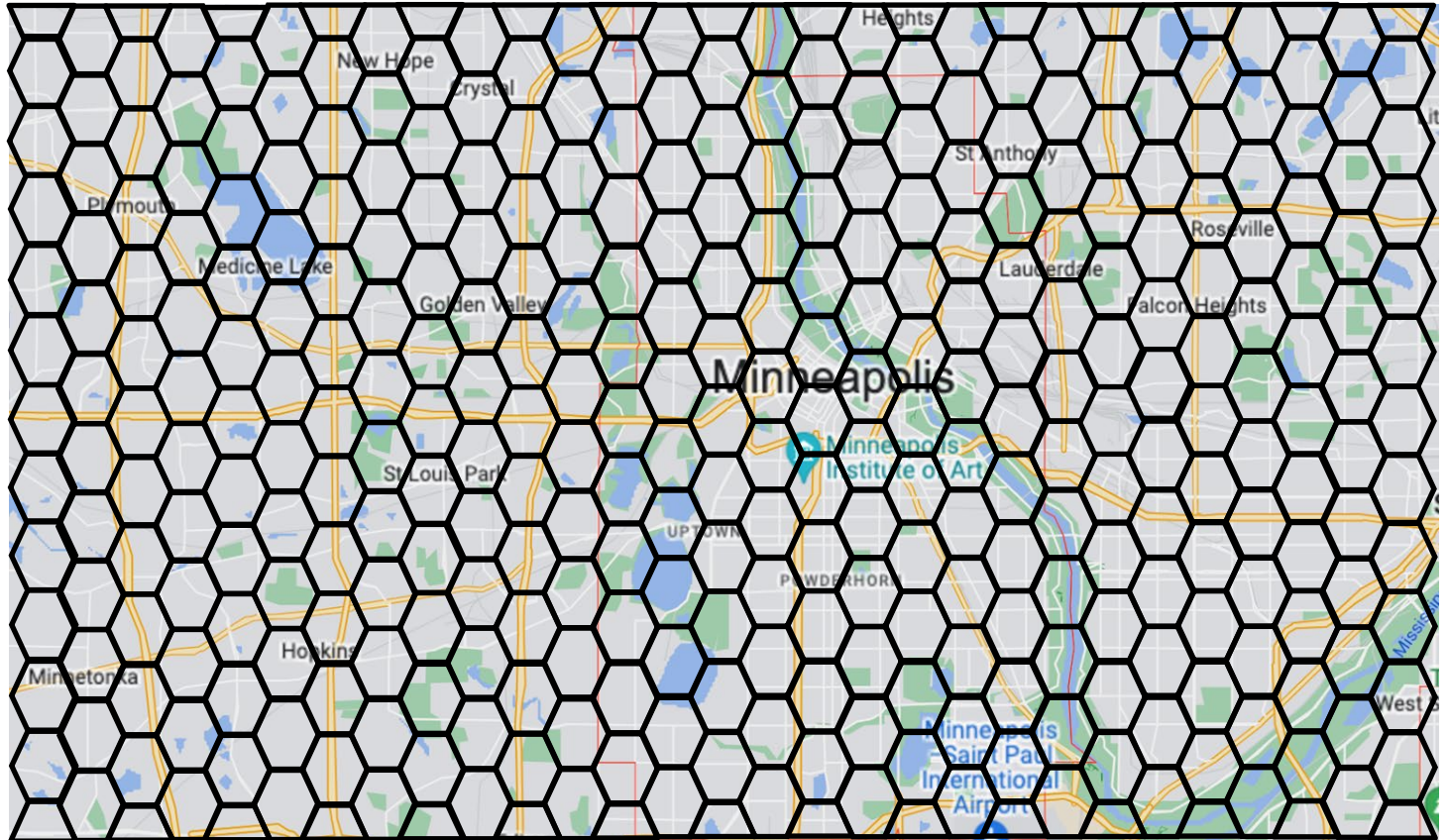
Random Sampling

Area of
interest



Random Sampling

Area of
interest



Random Sampling

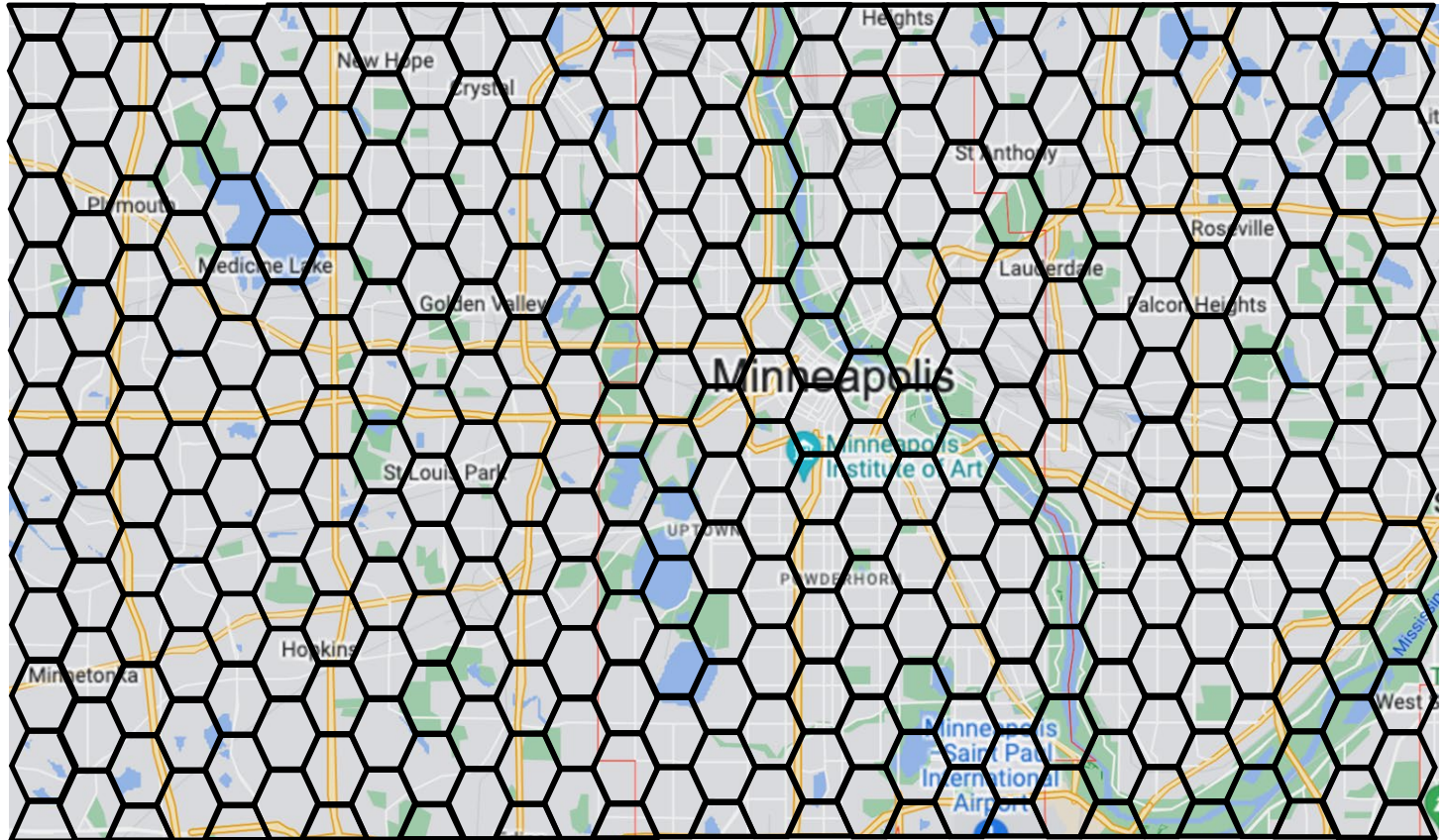
Monthly income, per capita (\$)

6,404

6,000

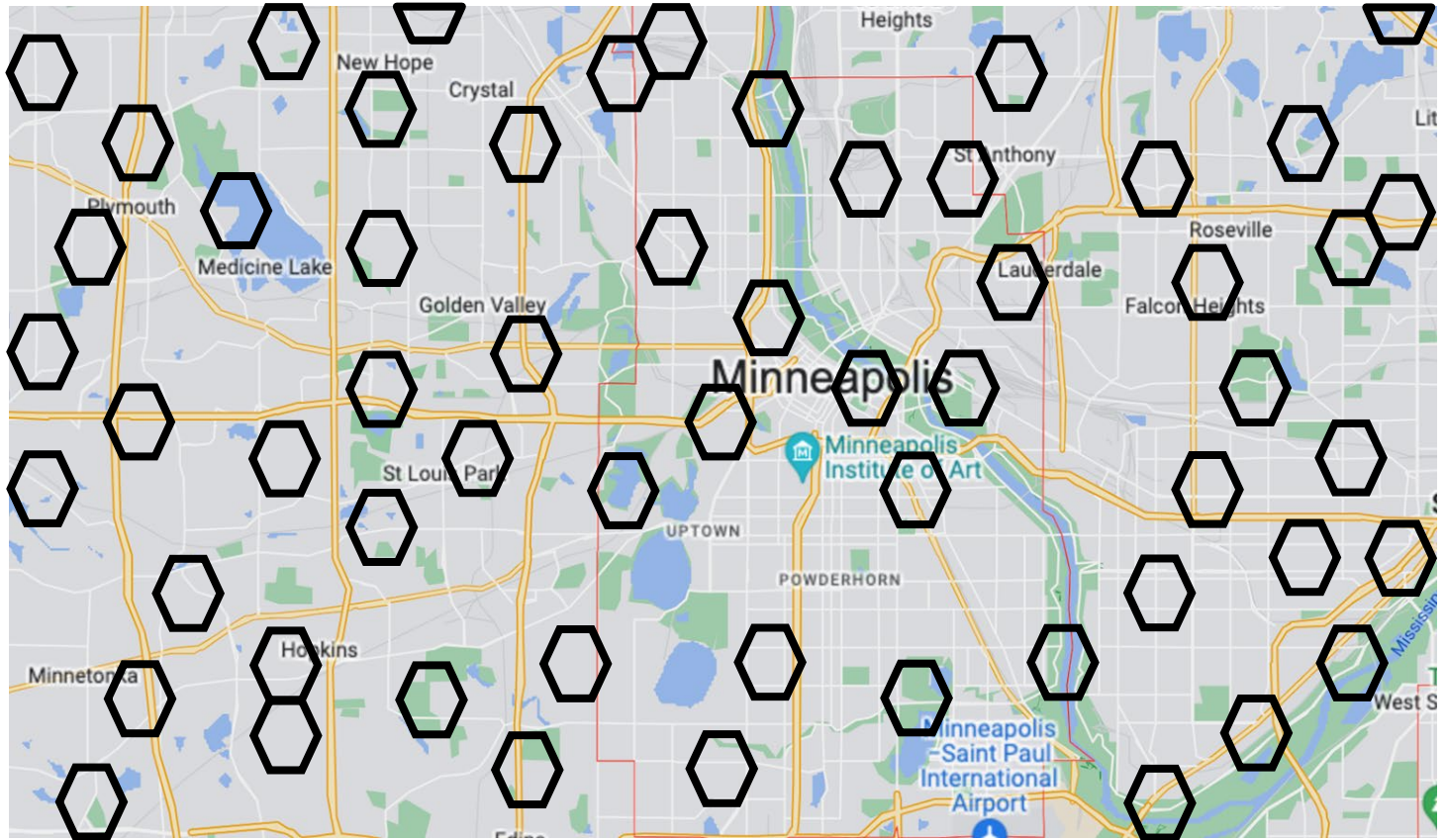
3,000

Population

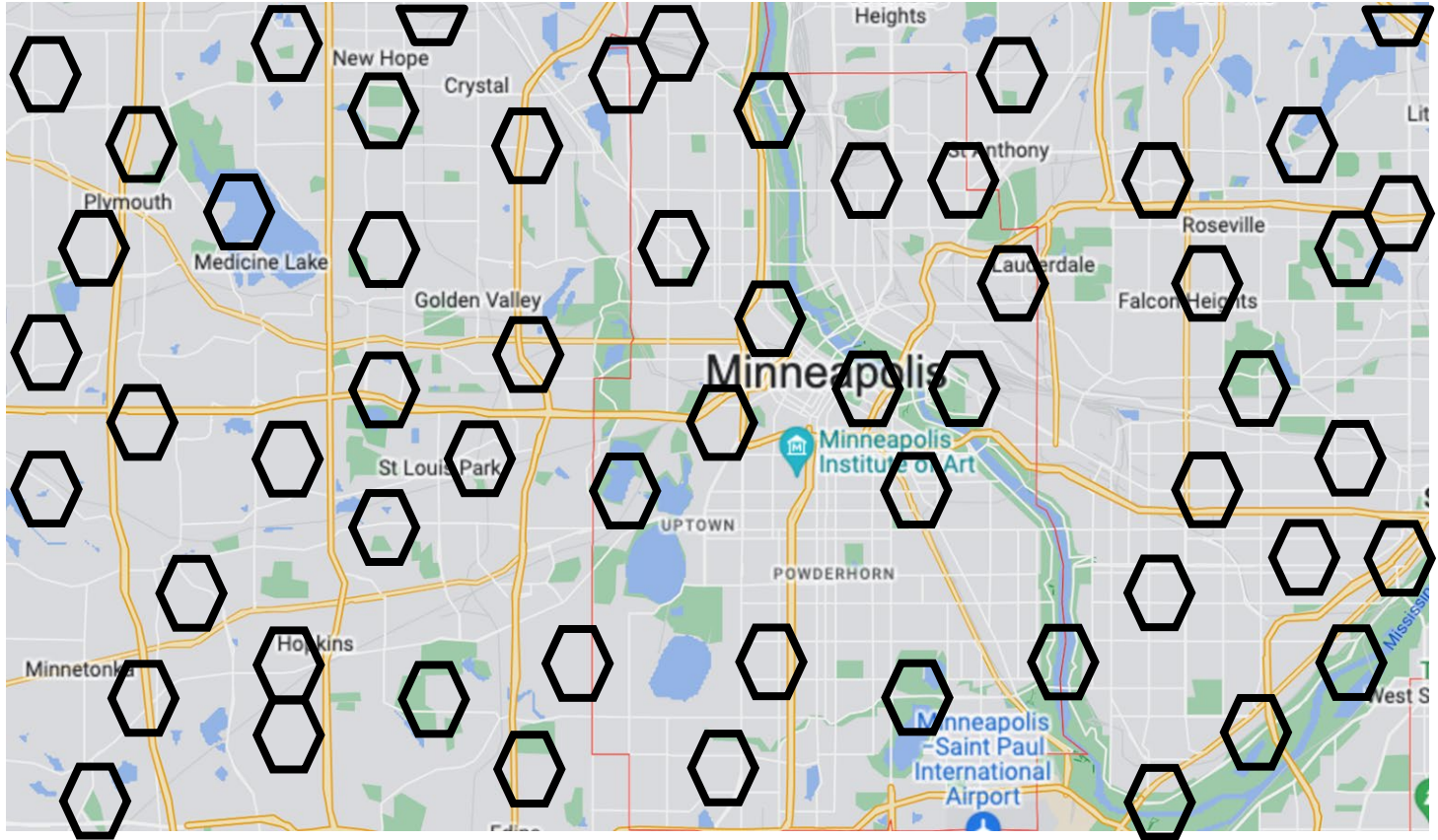
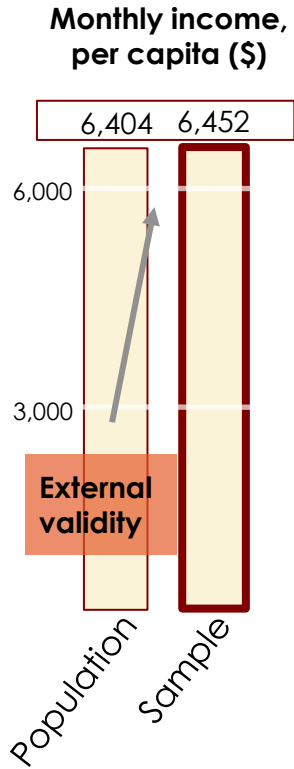


Random Sampling

Randomly
sample
from area of
interest



Random Sampling

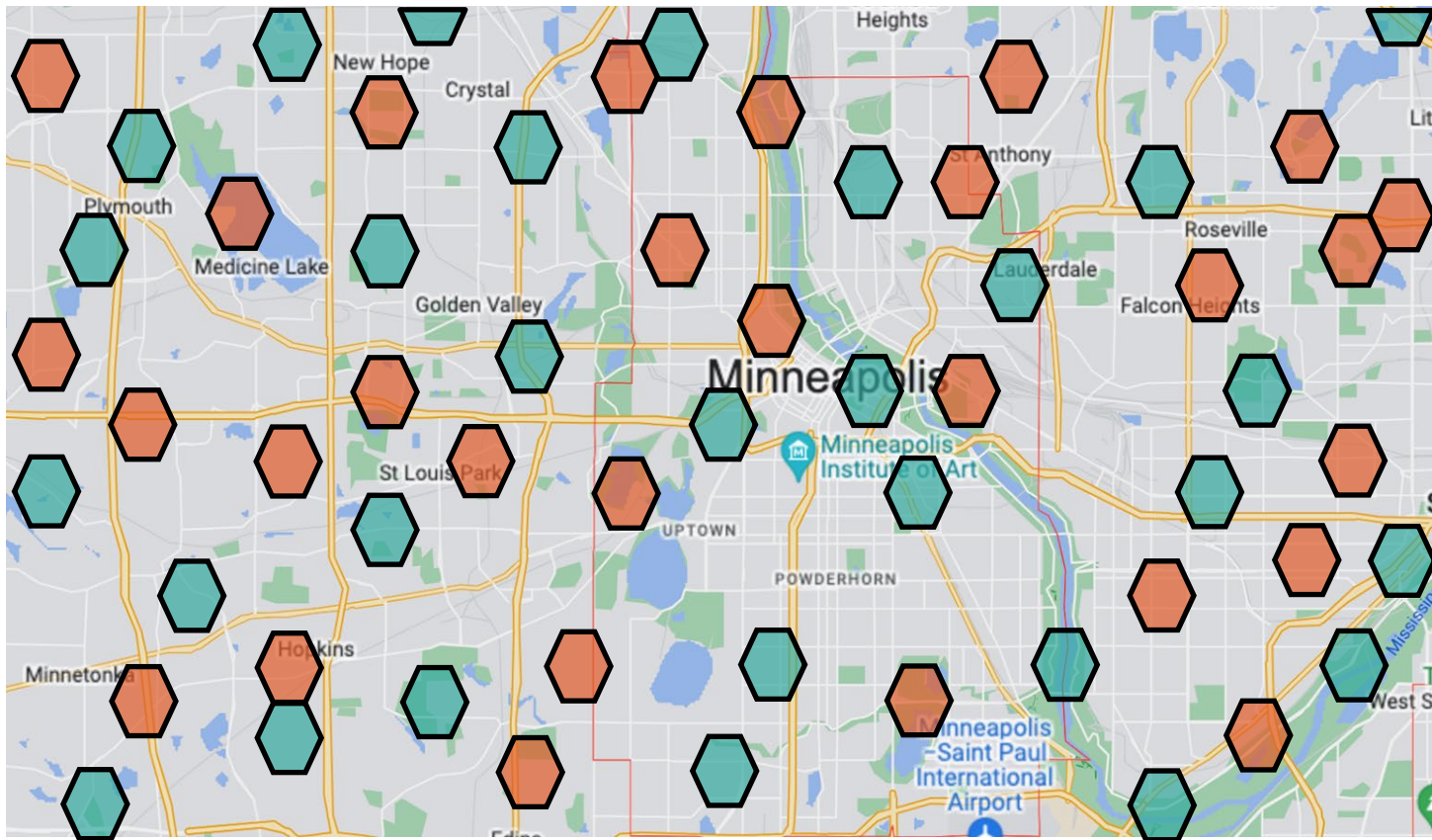


Lecture Overview

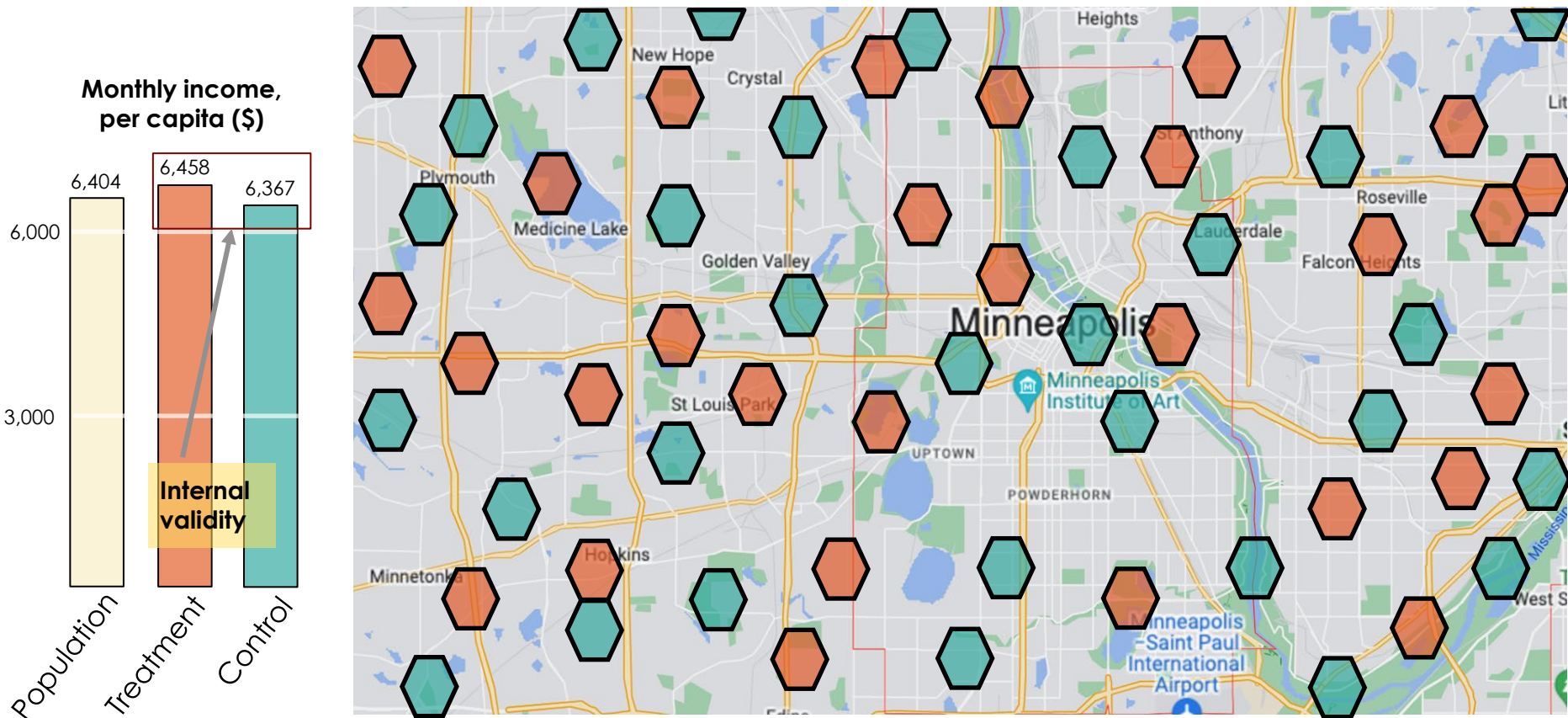
- What is Randomization?
 - Random Sampling
 - **Random Assignment**
- Randomization Procedures
- The Unit of Randomization
- Sample Size Considerations

Random Assignment

Randomly
assign to
treatment
and **control**



Random Assignment



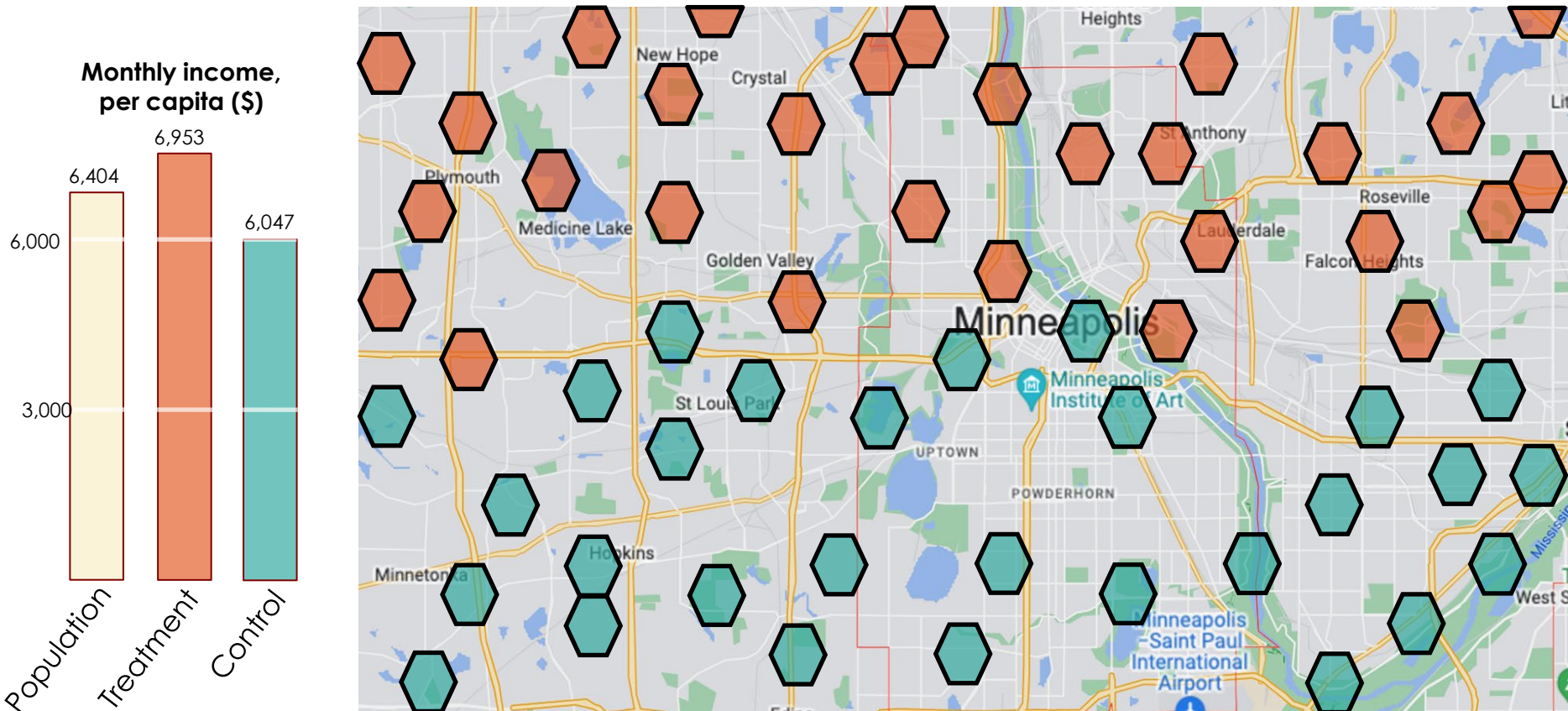
Suppose we take a random sample from our target population and assign the participants in the **northern half of the sample to the treatment**, and the participants in the **southern half of the sample to the control**. This is an example of:

- A. Random sampling but not random assignment
- B. Random assignment but not random sampling
- C. Random sampling and random assignment
- D. Neither random sampling nor random assignment
- E. Unsure

Suppose we take a random sample from our target population and assign the participants in the **northern half of the sample to the treatment**, and the participants in the **southern half of the sample to the control**. This is an example of:

- A. **Random sampling but not random assignment**
- B. Random assignment but not random sampling
- C. Random sampling and random assignment
- D. Neither random sampling nor random assignment
- E. Unsure

Not Random Assignment



Lecture Overview

- What is Randomization?
 - Random Sampling
 - Random Assignment
- **Randomization Procedures**
- The Unit of Randomization
- Sample Size Considerations

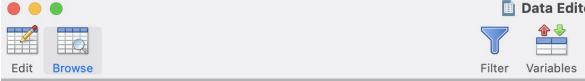
Basic Randomization Design

- Randomization can be appropriate when a program is oversubscribed and resource constraints prevent everyone who is eligible from receiving the treatment
- A “lottery” design can be easy to understand and implement
 - Randomly assign treatment and control from sample of those who are eligible
 - Already used in the real world
 - Examples: Old system of military recruitment was randomized and subsidized social housing units in Egypt.

Conducting a “Lottery” Design: Randomize with a Complete List of Study Participants

How does it work?

- Compile a list of all participants that will be in your study
- Determine the number of units that you want in treatment and in control
 - This number is based on logistical factors and on statistical power calculations
- Assign to treatment or control, typically using a random number generator. (ex: on random number generator on excel, stata, R)



	schoolid	areaid	pretest_mean	random	treatment
50	503	5	22.1532	.3928247	1
51	130	1	18.2115	.3937532	1
52	617	6	35.6716	.4257155	1
53	215	2	23.2667	.4363684	1
54	603	6	14.7627	.4451973	1
55	408	4	25.0333	.4523951	1
56	529	5	14.7895	.4550135	1
57	602	6	37.8	.4553497	1
58	633	6	29.506	.4629052	1
59	514	5	25.2937	.4742467	1
60	639	6	20.4862	.4871492	1
61	509	5	23.2927	.5233575	0
62	346	3	36.3158	.5371823	0
63	625	6	28.8974	.5402876	0
64	411	4	20.6154	.5585051	0
65	132	1	34.7232	.5694133	0
66	314	3	22.9275	.5766308	0
67	320	3	36.5556	.5860199	0
68	502	5	16.8955	.595137	0
69	114	1	16.3226	.5979754	0
70	202	2	19.5256	.6041152	0

Example: For each unit (e.g. a school), a random number generator has assigned a random number in the “random” column. Treatment is assigned to those with a number of less than 0.5

“Rolling Randomization Design”: Randomize Each Participant Upon Entry into Study

How does it work?

- Set the probability of assignment to treatment group to a fixed percentage (e.g., 50%, 75%)
- Conduct a point-of-service randomization
 - Example: when a new study participant comes to a study site, a research staff member could run randomization code through software to randomize the participant to treatment or control, after conducting a process of informed consent

ID	Coin	Treatment /Control
1	Heads	T
2	Heads	T
3	Tails	C
4	Heads	T
5	Tails	C
6	Heads	T
7	Tails	C
8	Tails	C
9	Heads	T
10	Heads	T
Count:		T: 6 C: 4



Multiple Treatments

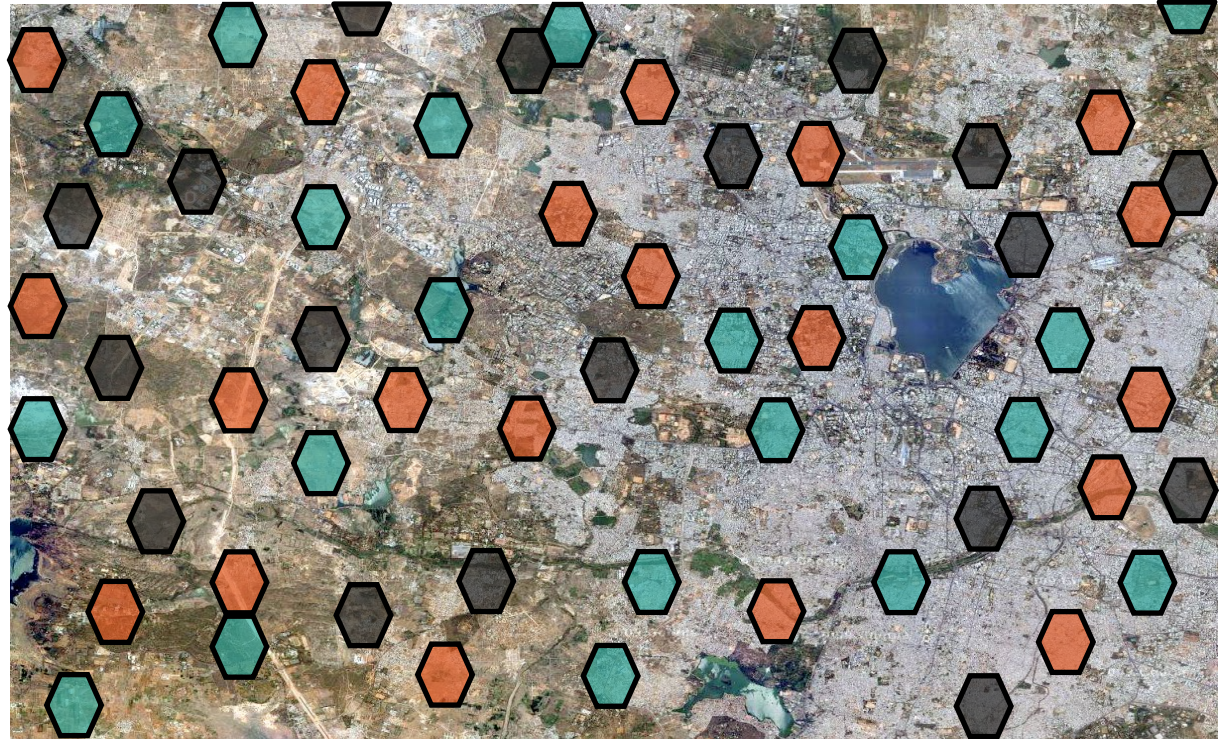
Question: How do different treatments compare?

Treatment 1

Treatment 2

Control

Pictured: Spandana, Hyderabad, the site of a microcredit study (Banerjee et al. 2015). These are not the actual treatment and control neighborhoods.



How do different groups compare within a study?

Example: Study A

Testing one intervention, compared to a comparison group that continues with “business as usual”

Intervention group:
Program intervention
Comparison group

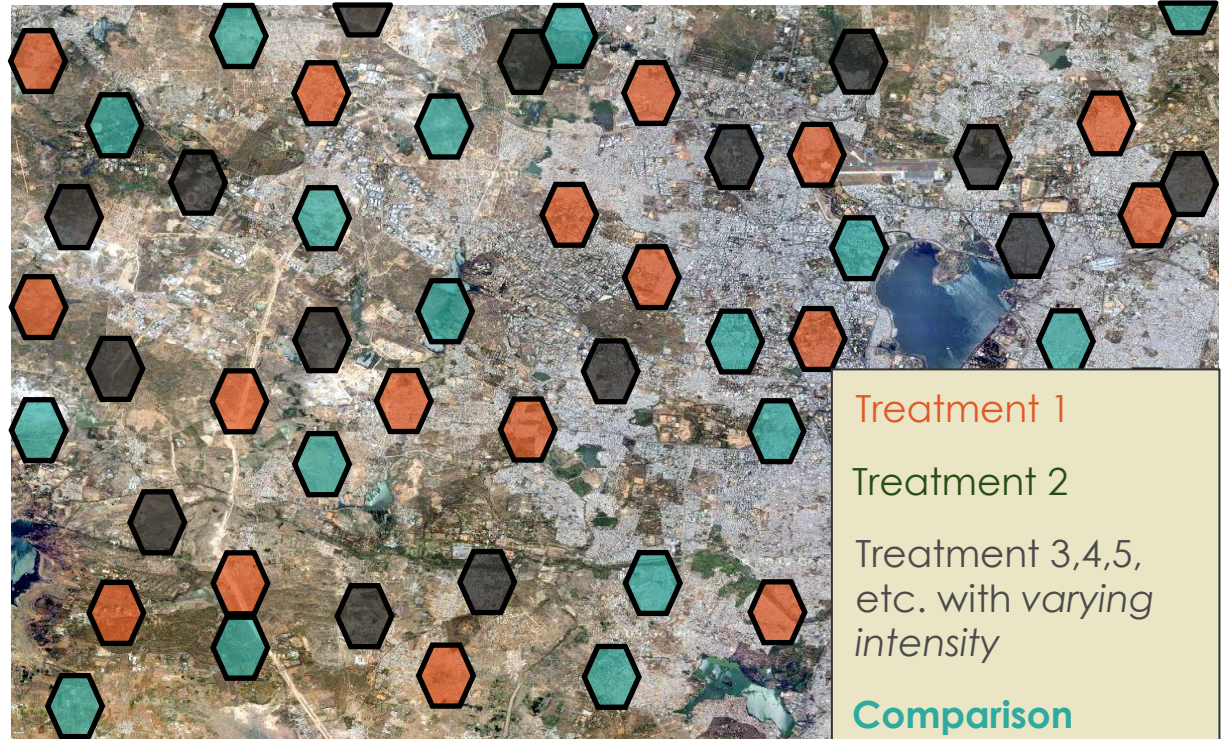
Example: Study B

Testing multiple interventions, compared to a control group that continues with “business as usual”

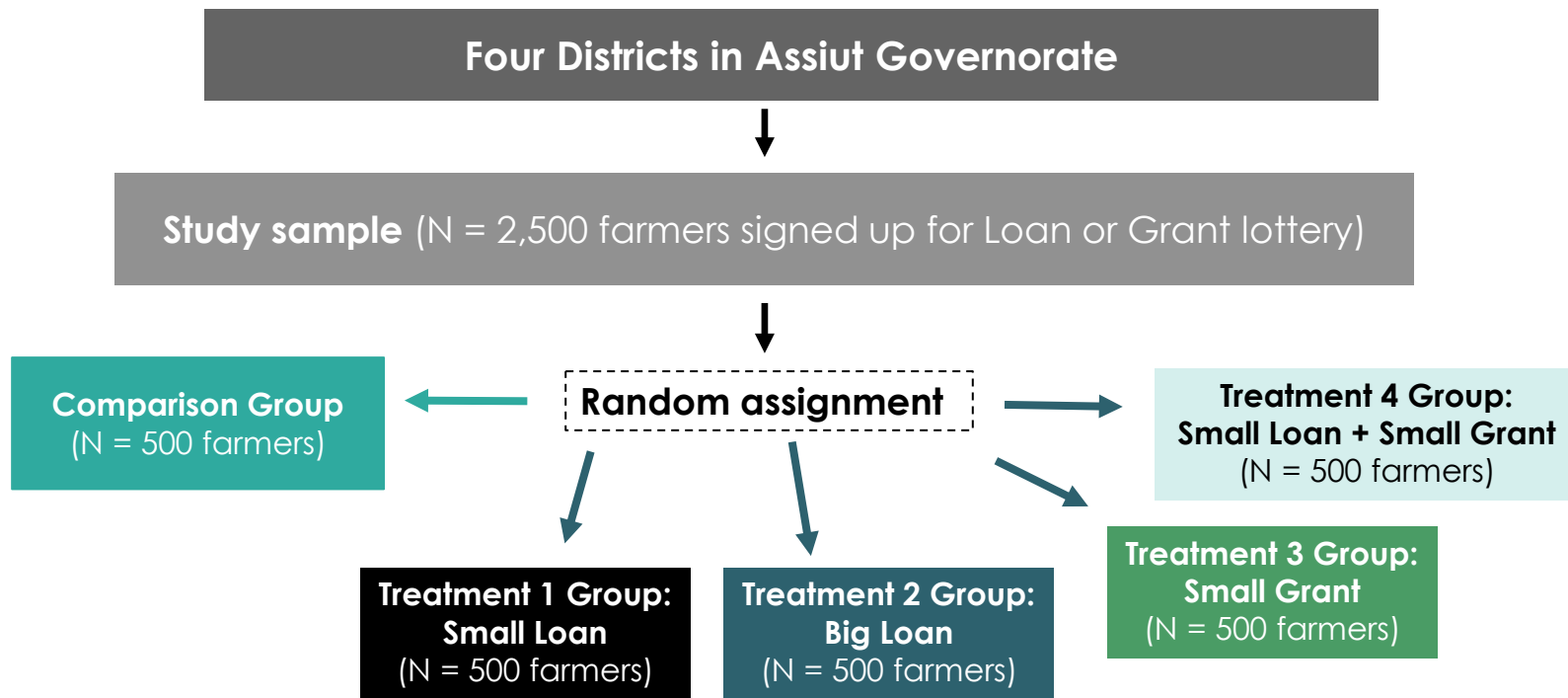
Intervention group 1:
Program intervention #1

Treatment groups:
Different program intervention
OR same intervention *with varying intensity*

Comparison group



Example: Micro-Finance Project



Relationship between question and the treatment groups

Question: What is the differential impact of the different treatments (does loan size and grant size matter for impact)?

Does the programs work **better together** than **separately**?

Lecture Overview

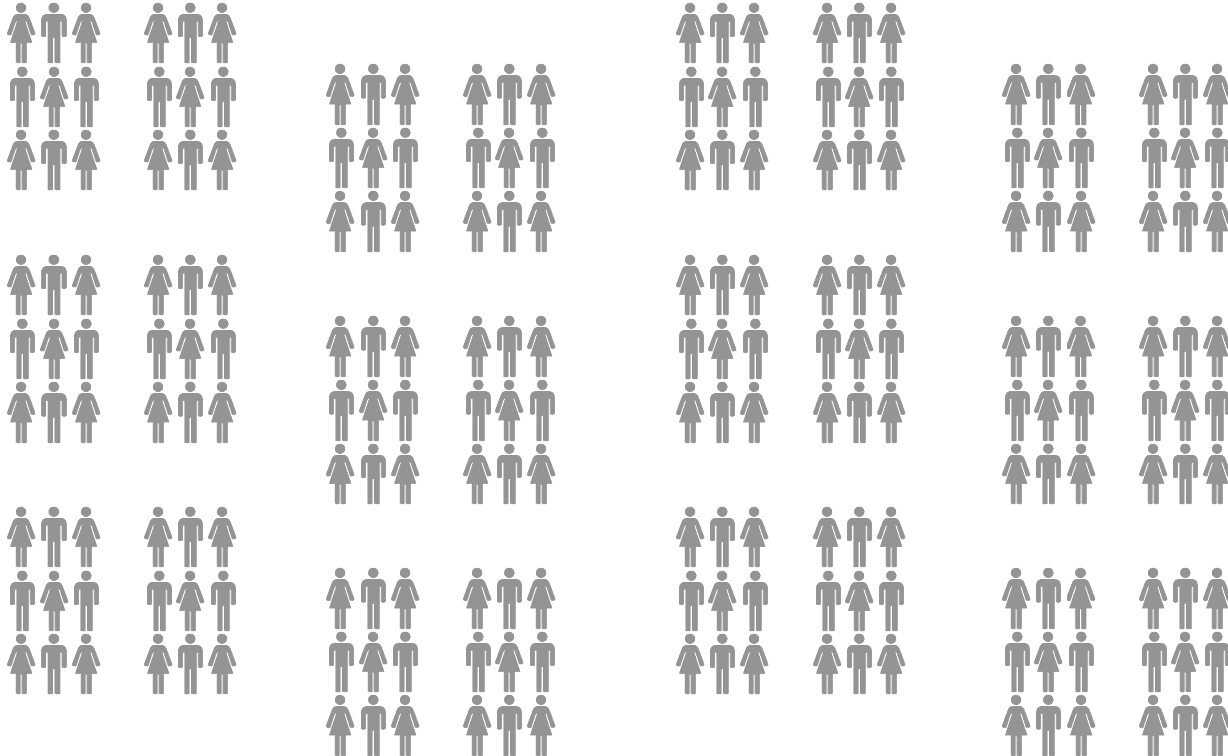
- What is Randomization?
 - Random Sampling
 - Random Assignment
- Randomization Procedures
- **The Unit of Randomization**
- Sample Size Considerations

Units of Observation and the Level of Randomization

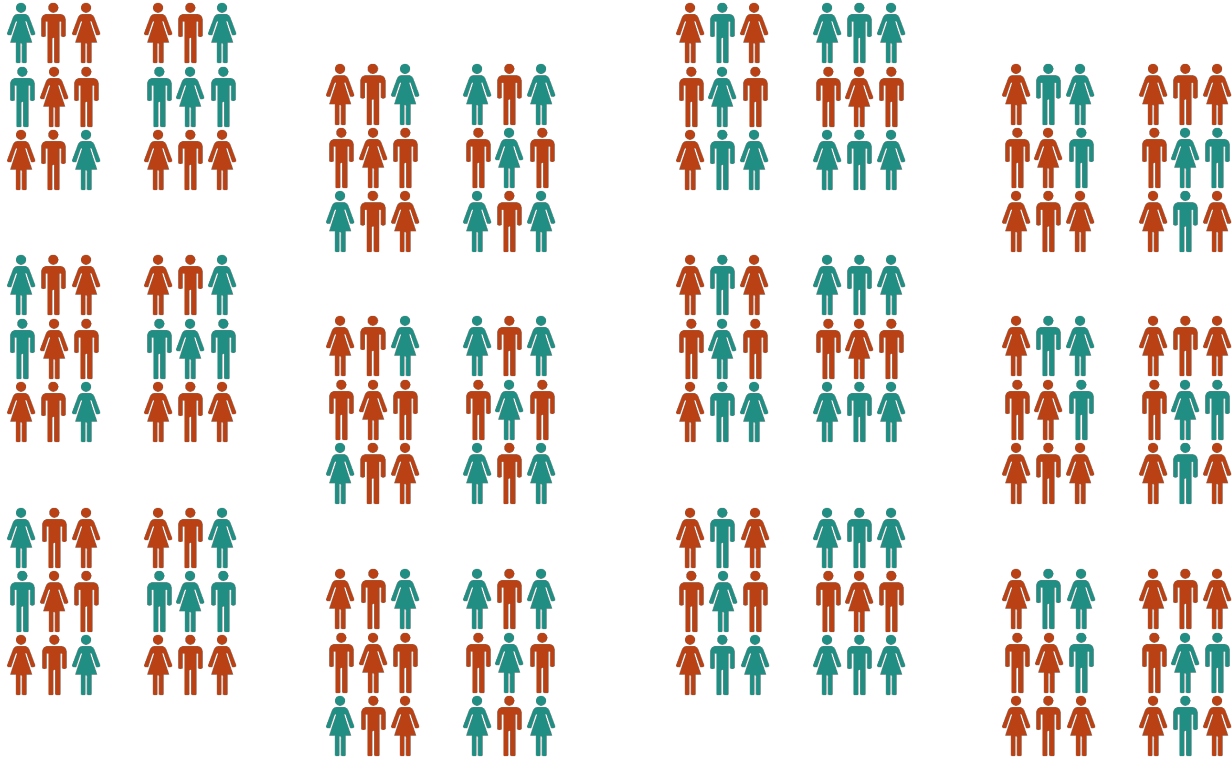
At which level should a study randomize?

- Randomizing at the individual level
 - e.g., people, or students
- Randomizing at the group level
 - e.g., villages or schools
 - *Outcomes can still be measured at the individual level*

Unit of Randomization: Individual?



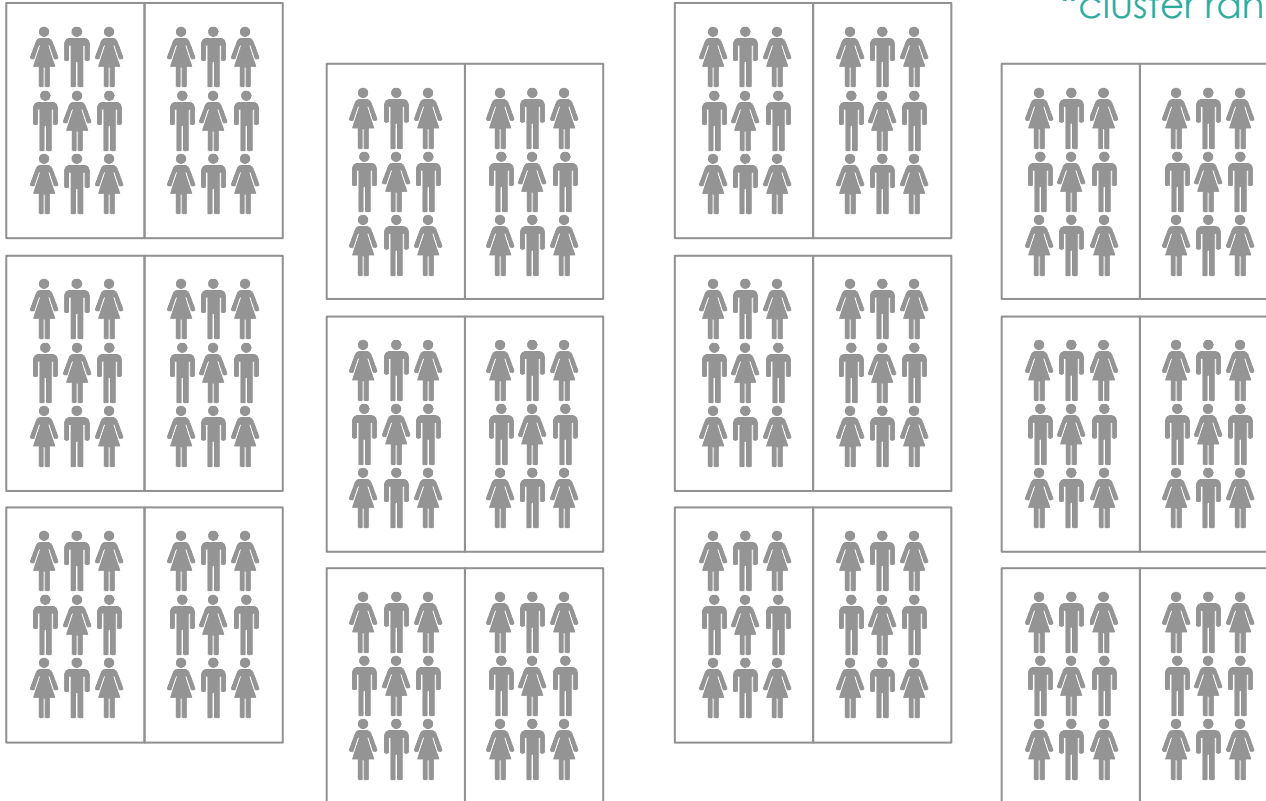
Unit of Randomization: Students?



Unit of Randomization: Classroom?

We call groups of units “**clusters**”
Randomization at the group
level:

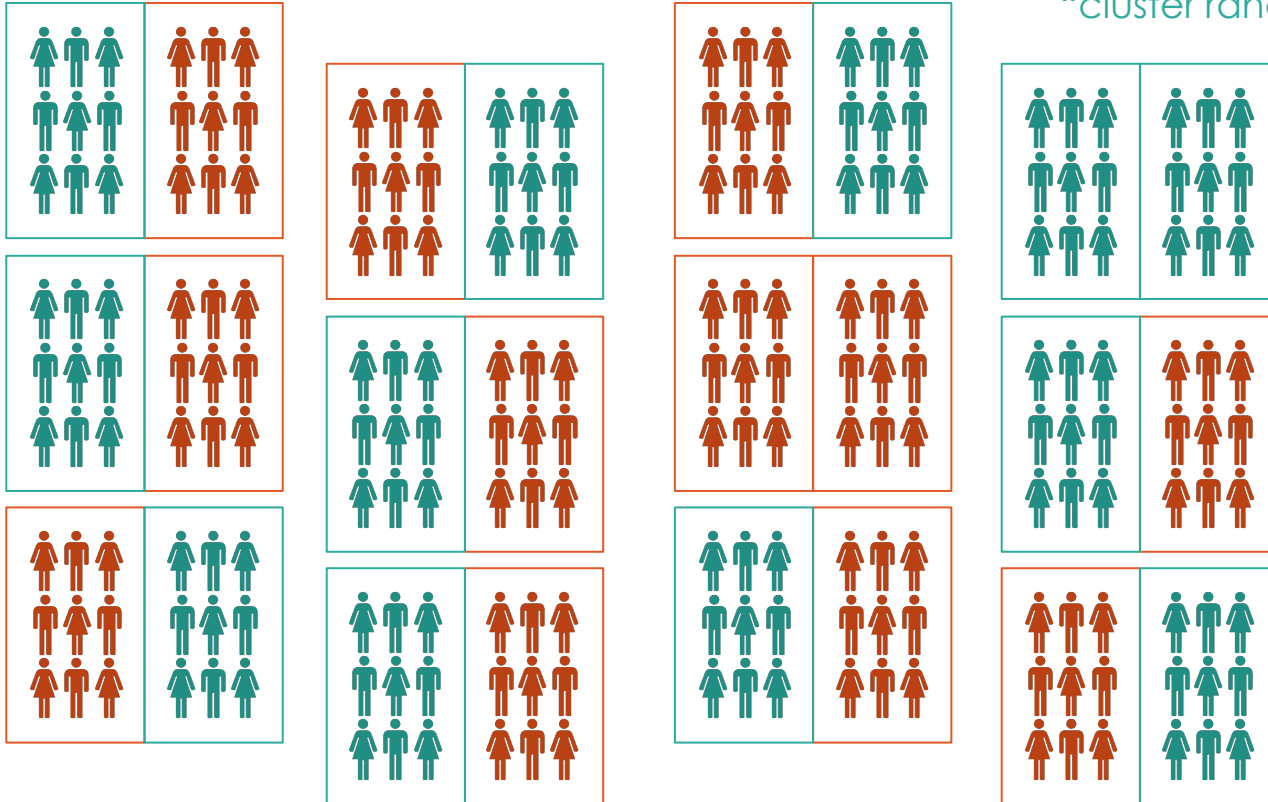
“cluster randomized trial”



Unit of Randomization: Classroom?

We call groups of units “**clusters**”
Randomization at the group level:

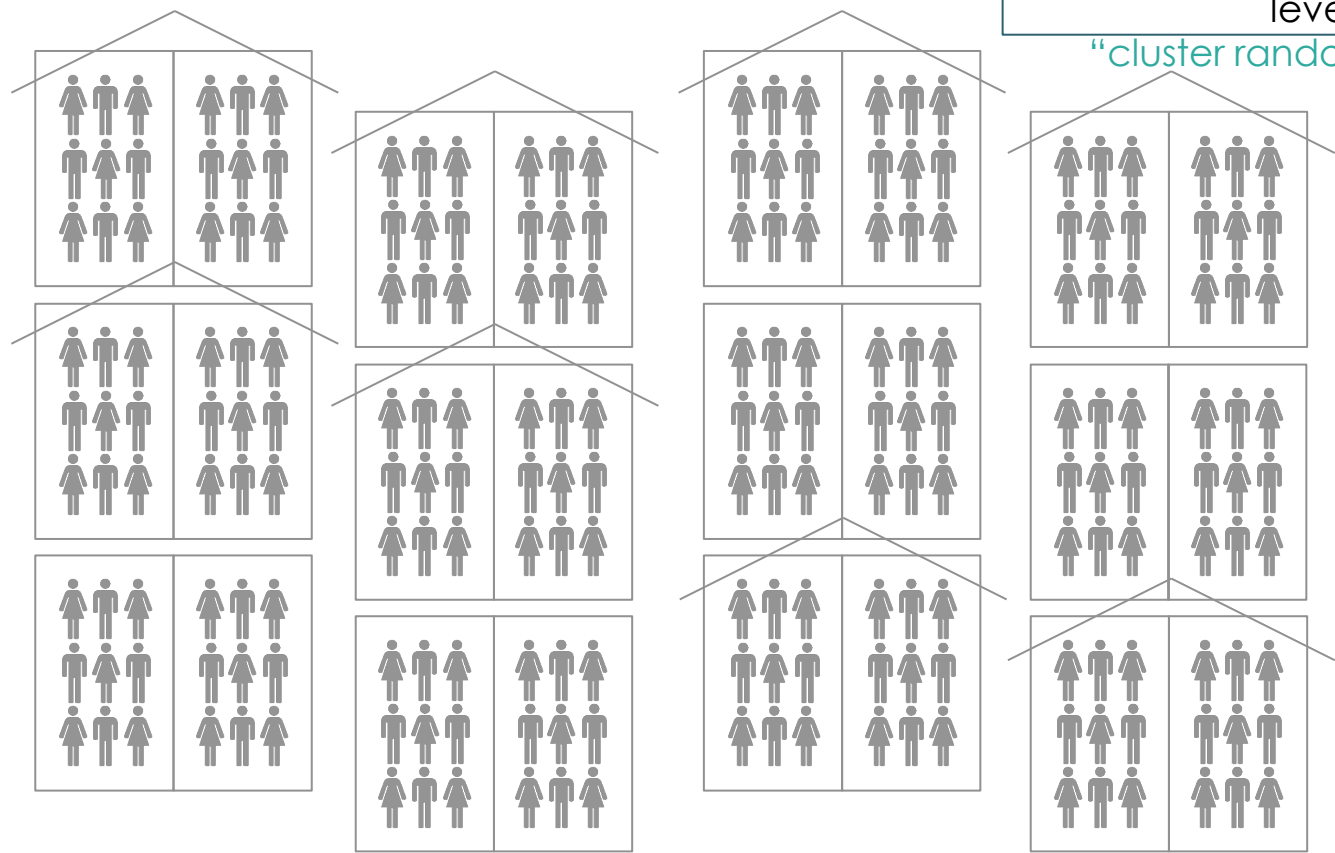
“cluster randomized trial”



Unit of Randomization: School?

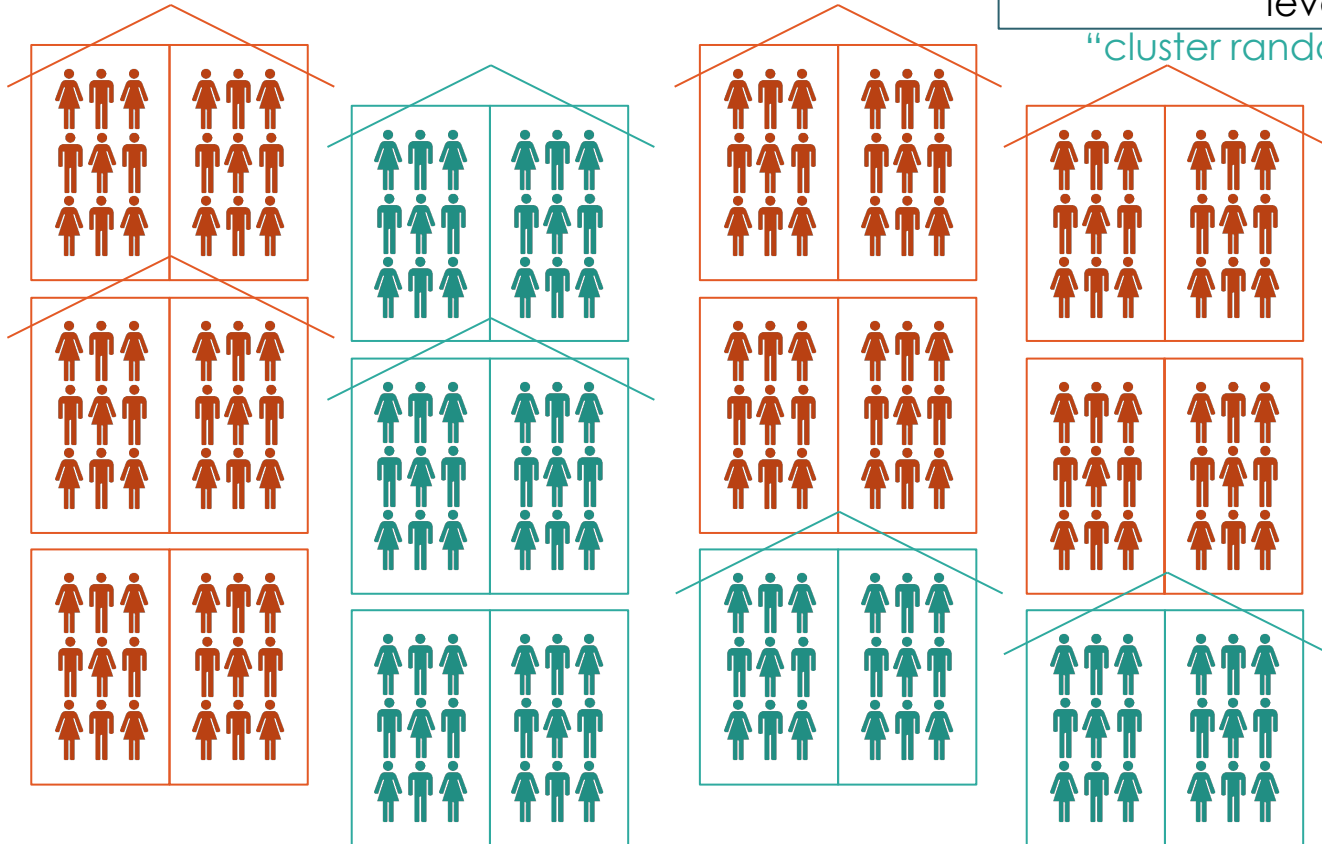
We call groups of units “**clusters**”
Randomization at the group level:

“cluster randomized trial”



Unit of Randomization: School?

We call groups of units “**clusters**”
Randomization at the group level:
“cluster randomized trial”



A state education department wants to see if increasing the duration of recess can help improve student's learning. **What is an appropriate unit of randomization, and why?**

- A. Student level
- B. Classroom level
- C. School level
- D. District level
- E. Unsure

A state education department wants to see if increasing the duration of recess can help improve student's learning. **What is an appropriate unit of randomization, and why?**

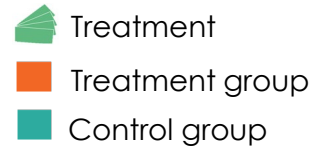
- A. Student level
- B. Classroom level
- C. **School level** – *There is no single correct answer, but in this scenario, randomizing at the school level might be the most feasible and appropriate option. If students in the same classroom or classes within the same school had different lengths of recess, this might cause conflict or confusion. Randomizing at the district level might not provide a large enough sample size.*
- D. District level
- E. Unsure

Level of Randomization: Considerations

Choose a level of randomization to minimize **noncompliance** and to measure or contain **spillovers** (these are two different concepts)



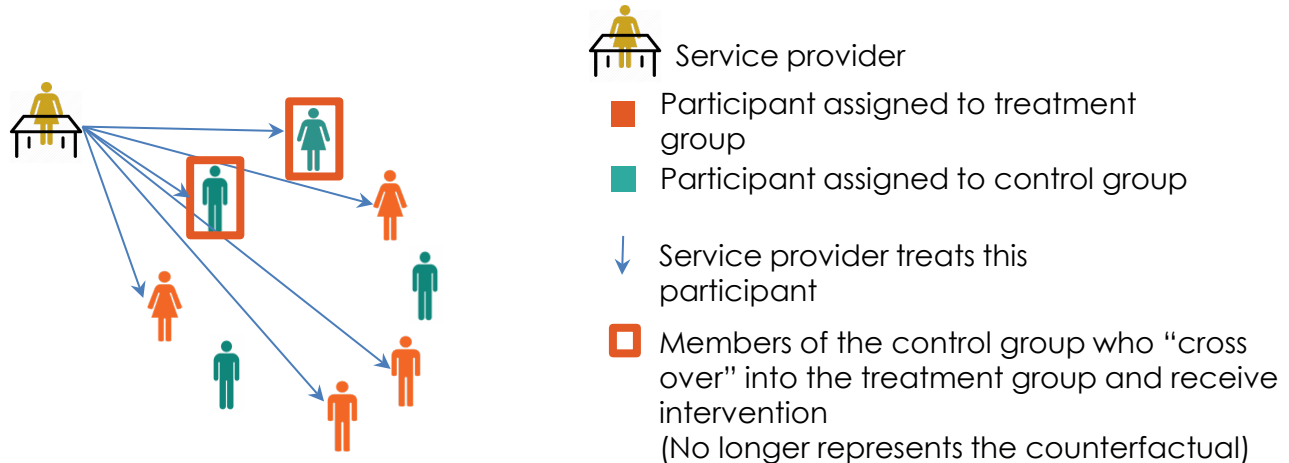
Noncompliance: When participants do not follow (“comply with”) their treatment assignment. For example, if participants assigned to the control group join the treatment group, or vice versa.



Spillover: When the treatment indirectly affects those who have not been treated. Spillover effects can be positive or negative.

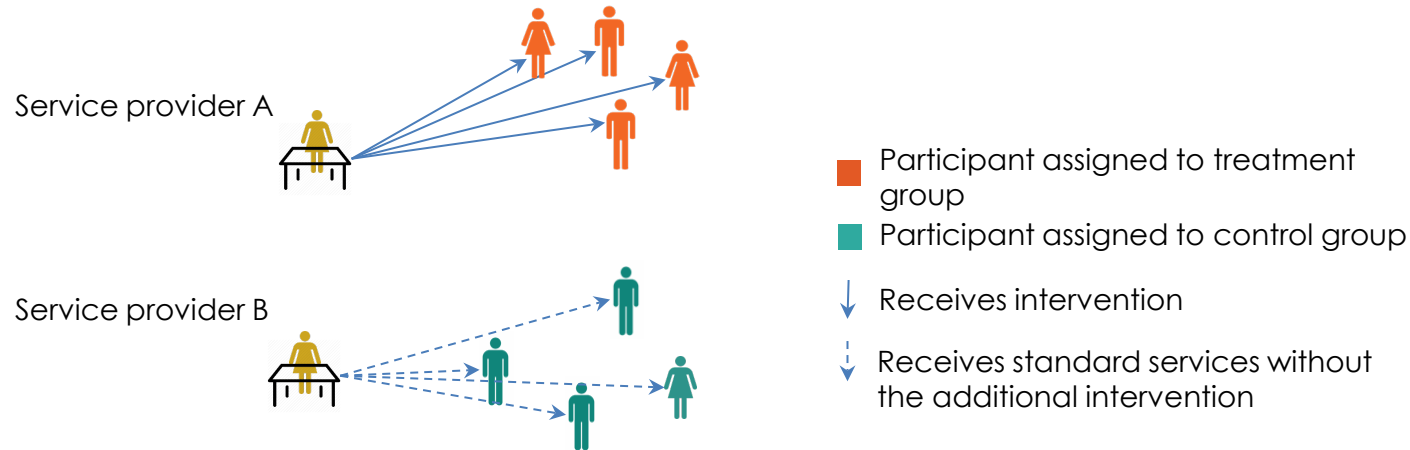
Potential Sources of Noncompliance

- Logistical or political challenges. For example, service providers may find it difficult to administer treatment alongside their other responsibilities.
- Service providers might have trouble distinguishing between treatment and control (or customizing service), or may be unwilling to provide differential treatment.



Solution: Change Unit of Randomization to a Different Level – Randomize Service Providers

- Have different service providers administer the different treatments
- Randomly assign treatment/control to those service providers
- This enables providers to treat entire clusters the same



Spillovers - Outcomes

Spillovers may not put a study in jeopardy if they are contained or measured, but problematic if they affect the control group

- Spillovers can be positive or negative
 - **Example of negative spillovers:** if the beneficiaries of a job matching program fill all available positions, this puts untreated job seekers at a disadvantage and makes it less likely that they will get a job.
 - **Example of positive spillovers:** production line workers whose coworkers receive soft skills training may benefit from the learnings of their coworkers, increasing their retention and overall productivity.
- Spillovers can cause impact to be underestimated or overestimated

More information: <https://www.povertyactionlab.org/resource/randomization>

Solution 1: Randomize at a Different Level

Randomizing at the production line level contains the positive spillovers within the treated production lines

Level of
randomization:
production line

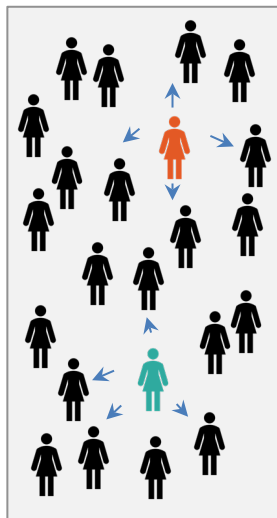


■ Treatment group
■ Control group

Solutions 2: Build in Buffers

Providing a buffer between the treatment and control subjects prevents treatment from spilling into control

Level of
randomization:
individual



- Treatment group
- Control group
- Not sampled
- Friends

Lecture Overview

- What is Randomization?
 - Random Sampling
 - Random Assignment
- Randomization Procedures
- The Unit of Randomization
- **Sample Size Considerations**

Why do we evaluate?

- Do we always want to scale a program that has shown positive effects?
- Do we always want to discard a program that shows no effects?

Evaluation Results Versus Underlying Truth

What we really
want to know



Reality/underlying truth

What we actually
measure/learn



**Evaluation
results**

Evaluation Results Versus Underlying Truth

Important note!

This is a thought experiment!
We can't measure "underlying truth"
so we are using our evaluation results
to approximate it.

What we really want to
know, but cannot observe

→ **Reality/underlying truth**

What we actually
measure/learn

→ **Evaluation
results**

No impact
detected

Impact
detected

No impact

Impact

No impact	Impact

Results Versus Underlying Truth

Let's use a recognizable example:

- Imagine that tomorrow you wake up with a cough and sore throat.
- At first you think it's allergies, but realizing you are about to attend an indoor gathering with your entire family, what do you do next?

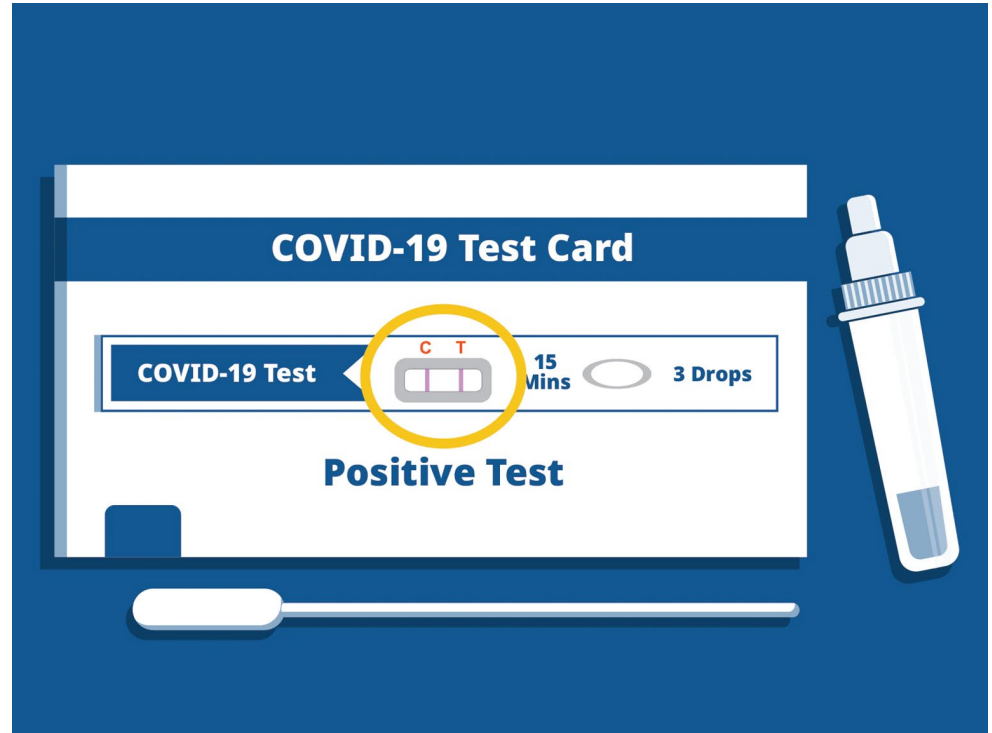


Image: mass.gov

Results Versus Underlying Truth

Rapid test results:
Do you test positive or negative?

Negative test results

Positive test results

Reality/underlying truth:
Are you infected with Covid-19?

Not sick

Sick

Results Versus Underlying Truth

Reality/underlying truth:
Are you infected with Covid-19?

Not sick

Sick

Rapid test results: Do you
test positive or negative?

Negative
test results

**Accurate
/true
negative**

False negative:
you conclude
you are NOT sick
when you are.

Positive test
results

False positive:
you conclude
that you are sick
when you are
not!

**Accurate
/true
positive**

How does this relate to RCTs and impact evaluation?

- Impact evaluation/RCTs are not *exactly* the same as a rapid Covid test! (of course!)
- But some of the underlying logic is similar:
 - We can't observe the underlying reality/underlying truth
 - We **take a sample** and use that sample to try to learn something about the underlying truth.
 - And we want to minimize false positives and false negatives to the extent possible!

Evaluation Results Versus Underlying Truth

		Reality/underlying truth	
		No impact	Impact
Evaluation results	No impact detected	GREAT!	Mismatch!
	Impact detected	Mismatch!	GREAT!

Evaluation Results Versus Underlying Truth

		Reality/underlying truth	
		No impact	Impact
Evaluation results	No impact detected	GREAT!	False negative: you conclude there is NO impact when there is.
	Impact detected	False positive: you conclude there is impact when there is not.	GREAT!

What are some risks if you find a **false positive** (finding an “impact” when there isn’t actually an impact)?

What are some risks if you find a **false negative** (finding no impact when there actually is an impact)?

What is statistical power?

The **statistical power** of an evaluation reflects how likely we are to detect a given change in an outcome of interest.

In other words, statistical power is the likelihood of **avoiding a false negative** (concluding there is no impact when there actually is one).

By convention, we aim for 80% power.

- This means that we expect that 80% of the time we will be able to detect an effect if there is one.
- 20% of the time we will falsely conclude there is no impact of our program.

Statistical Significance

What do we mean when we say “Detect an effect”?

→ Find a **statistically significant** impact

Statistical significance is about **avoiding a false positive** (concluding your program had an impact when it did not).

We want to be sure that the measured program effect is due to the program itself and not due to natural variation or random chance.

When we do our analysis, we ask ourselves:

- What is the probability that we would observe this outcome/measure this difference between treatment and control groups if in fact our program had **no** impact?

Statistical Significance & Statistical Power

Statistical significance

- “Detecting an effect” = measuring a statistically significant impact.
- Avoiding false positives a.k.a. falsely concluding there was an impact when there is none.
- By convention, we set statistical significance to 95% or higher.

Statistical power

- Our likelihood of detecting an effect if there actually is one!
- Avoiding false negatives a.k.a. falsely concluding there is no impact.
- By convention, we aim for 80% power.
 - This means that we expect that 20% of the time we will falsely conclude there is no impact of our program.

Power: Main Ingredients

1. **Sample size:** a larger sample means that treatment and control are more representative of the overall population, making it easier to distinguish an effect.
1. **Effect size:** a large effect is easier to distinguish from zero than a small effect.
1. **Variance:** lower variability in the outcome variable makes it easier to distinguish an effect.
1. **Sample split:** an equal proportion in treatment and control makes it easiest to distinguish an effect.

Sample Size

- **What:** The number of individuals or units in your evaluation.
 - Remember to account for attrition (individuals dropping out of the evaluation/program)!
- **Example:** A health intervention aims to reduce smoking rates in Alexandria. If fewer smokers than expected agree to participate (low take-up),
 - Consider attrition: 300 people may enroll, but only 100 persist through the program. Attrition can be either someone dropping out of your program or missing from data collection.
- **Think about:** How many individuals does the program serve? How many will be in treatment vs. control? Over how long?

This is the input to power that you have the most control over and will thus often be your main lever to maximize power.

How does sample size interact with power?

Two main takeaways here

(relationship between power, effect size, sample size):

- For a given effect size: higher sample size will yield more statistical power.
- If statistical power is set: higher sample size means you can detect a smaller effect.

Effect size & minimum detectable effect (MDE)

- **What:** The impact your program has on the outcome(s) of interest.
 - Remember to account for the the number of people who take up/participate in the treatment.
- **Example:** A financial literacy program might lead to a \$50 increase (on average) in savings for the treated group.
 - Consider take-up: If you offer the seminar to 300 people and only 150 show up, all 300 would still be in your treatment group!
- **Think about** what is meaningful in your context! Want to be powered for the **smallest effect size that is still meaningful.**

MDE size

Minimal Detectable Effect (MDE)

The minimal effect size that can be detected with given statistical power (probability of correct positive, e.g. 80%), statistical significance (probability of a false positive, e.g. 5%) and sample size N.

- Ask: is it reasonable to expect effects as large or larger than the MDE?
- Would I like to be able to detect effects smaller than the MDE?
- Based on the MDE, can **adjust the sample size** to get to a realistic experimental design.

Variation in the outcome data

- **What:** How similar are individuals in your sample to each other (on the outcome of interest)?
- **Example:** In assessing the impact of an irrigation project on farm outputs across different governorates in Egypt, typically you'll find high variability in soil quality and farming practices in the different governorates and within each one, between different lands.
- **Where to find it?**
 - Program data (previous historical data are a good source of information).

Why does variation matter?

This relates to the earlier discussion on statistical significance and natural variation!

Before intervention

TREATMENT
GROUP



CONTROL
GROUP



Before the intervention, the treatment and control groups have exactly the same savings levels and **low variation** across participants.

Proportion of the sample in the treatment group

- **What:** How the sample size is split up between the program group and the comparison group (for example: 50% treatment and 50% comparison; 65% treatment and 35% comparison)
- **How to determine:** It depends! This is determined by program and resource constraints (for example, a given minimum or maximum number of participants, or budget constraints) and by other statistical power considerations (for example, the minimum effect size you aim to detect).

Power is maximized when the sample is equally split between treatment and control groups.

- For a fixed sample size, power maximized with even split between treatment and control.
- However, **power increases when the sample size increases** even if that comes through adding individuals just to one group (or unequally.)
 - Even you are unable to add more people to the treatment group due to resource constraints, you **could add more individuals to the control group**

Recap: Rules of thumb for statistical power

All else equal:

- \uparrow sample size = \uparrow power
 - \uparrow attrition = \downarrow power
- \downarrow expected effect = \downarrow power = \uparrow sample size needed
 - \downarrow take-up = \uparrow sample size needed
- \uparrow variation in outcomes = \uparrow sample size needed
- Power maximized when sample evenly split between T and C
- When individuals within clusters are similar \Rightarrow \uparrow sample size needed



Thank you