UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

The Structure and Design of Randomized Control Trials (RCTs)

illinois.edu

Outline for the Session

1. What are field experiments?
2. Why randomize?
3. How do I incorporate randomized evaluations into my research design?
4. What are the practical design and implementation issues?

What are Field Experiments?

(Recent) History

- Two worlds
  - Lab experiment research world
    - *Trades off control for context*
  - Observational research world
    - *Frustrated with identification challenge*

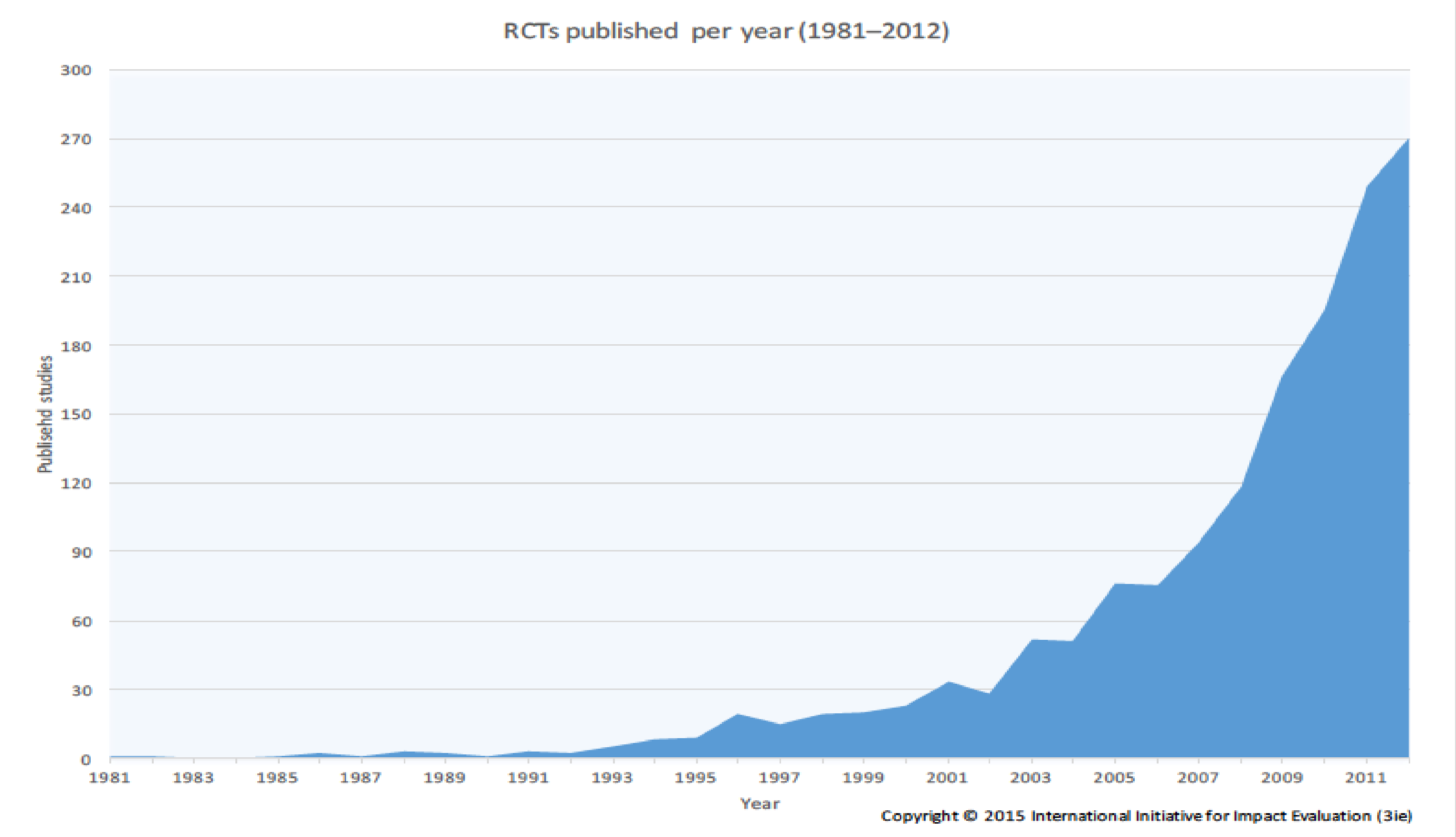
Broad Categorization

- Randomized evaluations
  - Aka randomized control trials (RCTs)
  - Key variation: What do participants know about the study?
    - *Fully unaware?*
    - *Unaware of randomization, aware of measurement (most development studies)?*
    - *Fully (or mostly) aware of randomization and measurement?*

Broad Categorization

- Lab experiments in the field
  - Aka framed field experiments or survey experiments
  - (sometimes) Aka incentive compatible surveys
  - Key variation:
    - *Outcome measure for larger study?*
    - *Full study itself?*

(Recent) History: Development



Copyright © 2015 International Initiative for Impact Evaluation (3ie)

Why Randomize?

The Problem of Causal Inference

- The potential outcome (Rubin, 1974)
- Average effect

$$E[\delta] = E\left[Y_i^T - Y_i^C\right]$$

The Problem of Causal Inference

- The potential outcome (Rubin, 1974)
- Treatment effect

$$E[\delta] = E\left[Y_i^T \middle| \mathrm{T}\right] - \mathrm{E}\left[Y_i^C \middle| \mathrm{C}\right]$$

The Problem of Causal Inference

- The potential outcome (Rubin, 1974)

$$
\begin{aligned}
E[\delta] = E\left[Y_i^T \middle| \mathrm{T}\right] - \mathrm{E}\left[Y_i^C \middle| \mathrm{C}\right] \\
-\mathrm{E}\left[Y_i^C \middle| \mathrm{T}\right] + \mathrm{E}\left[Y_i^C \middle| \mathrm{T}\right]
\end{aligned}
$$

The Problem of Causal Inference

- The potential outcome (Rubin, 1974)

$$\begin{aligned} E[\delta] = {} & E\left[Y_i^T \middle| \mathrm{T}\right] - \mathrm{E}\left[Y_i^C \middle| \mathrm{C}\right] \\ & -\mathrm{E}\left[Y_i^C \middle| \mathrm{T}\right] + \mathrm{E}\left[Y_i^C \middle| \mathrm{T}\right] \end{aligned}$$

$$= E\left[Y_i^T - Y_i^C \middle| \mathrm{T}\right] + \mathrm{E}\left[Y_i^C \middle| \mathrm{T}\right] - \mathrm{E}\left[Y_i^C \middle| \mathrm{C}\right]$$

The Problem of Causal Inference

- The potential outcome (Rubin, 1974)

$$\begin{aligned} E[\delta] &= E\left[Y_i^T \middle| \mathrm{T}\right] - \mathrm{E}\left[Y_i^C \middle| \mathrm{C}\right] \\ &\quad - \mathrm{E}\left[Y_i^C \middle| \mathrm{T}\right] + \mathrm{E}\left[Y_i^C \middle| \mathrm{T}\right] \end{aligned}$$

$$= \underbrace{E\left[Y_i^T - Y_i^C \middle| \mathrm{T}\right]}_{\text{Treatment Effect}} + \underbrace{\mathrm{E}\left[Y_i^C \middle| \mathrm{T}\right] - \mathrm{E}\left[Y_i^C \middle| \mathrm{C}\right]}_{\text{Selection Bias}}$$

Randomization Solves the Selection Bias

- First randomly select sample N from population P
- Second, randomly assign N into
  - Treatment (N_T) and Control (N_C)
- Since treatment is randomly assigned selection bias is removed
  - $E[Y_i^C|T] - E[Y_i^C|C] = 0$
- Then we can simply run the regression
  - $Y_i = \alpha + \beta T_i + \epsilon_i$
  - However, the SE are not generally correct if group variances differ

Caveats

- This requires two assumptions
  - SUTVA (Stable Unit Treatment Value Assumption)
    - *"no spillovers"*
  - Unconfoundedness/Ignorability
    - *"assignment to treatment is independent of outcome"*
- In most cases only partial randomization occurs
  - Population of study is not nationally representative but chosen conditional on some observables (poverty, age, gender, etc.)

Incorporating Randomized Evaluations in a Research Design

Preparing to Run a Field Experiment

1. Use economic theory to guide your design
2. Understand the local context
3. Obtain sufficient sample size

1. Use Economic Theory to Guide Your Design

- Theory allows appropriate nulls to be tested, designs to be efficient, and the ‘whys’ to be answered
- Theory is portable, many empirical results are not

An Example

- Go beyond A/B experiments to test economic theory
- List, 2004
  - Why do people receive different price quotes for the same good?
  - Economists have two major theories
    - ***Discrimination***
    - ***Search Costs***

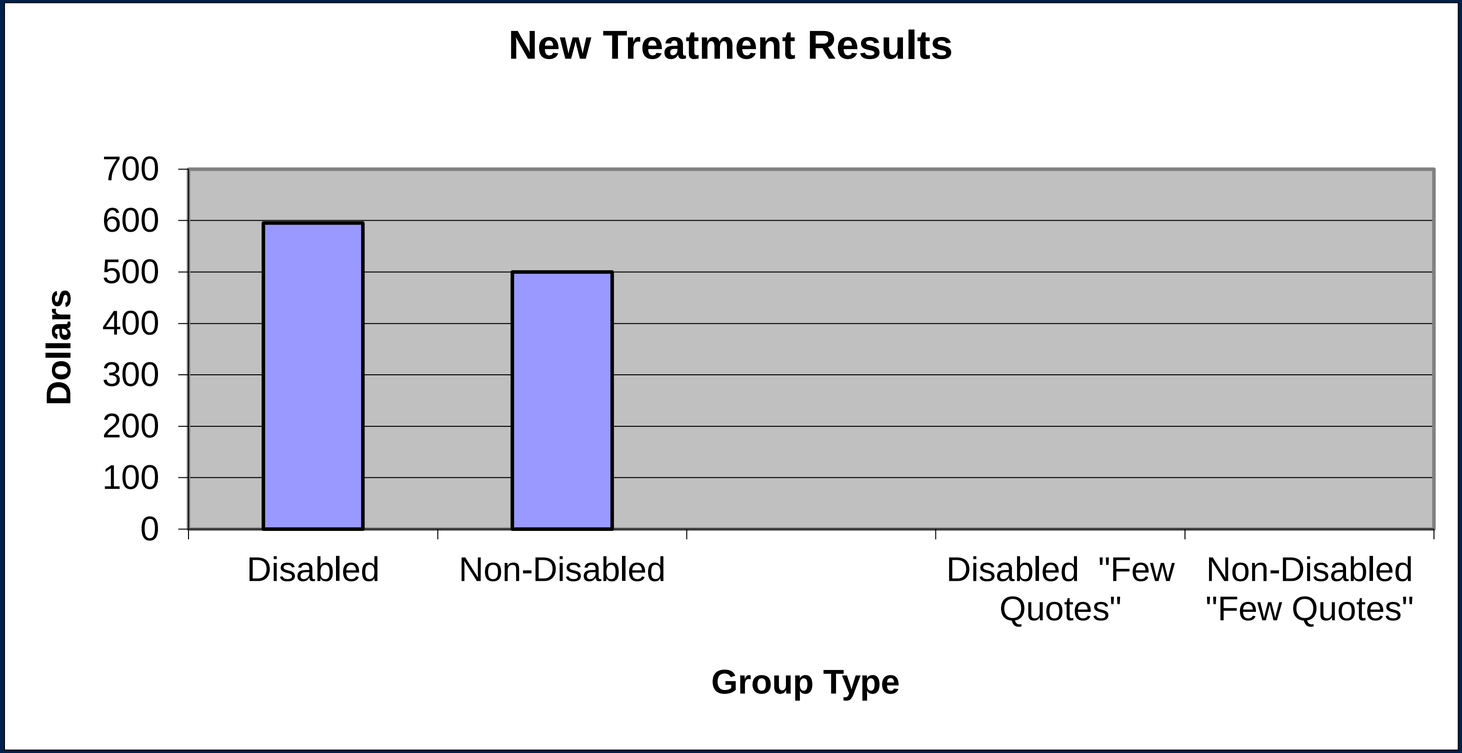
Discrimination NFE

- 12 disabled and 12 non-disabled testers approached various body shops in Chicago with different cars (identical cars across disabled and abled) that were in need of repair
- Offer differences:
    - Disabled receive prices 30% higher than the non-disabled receive

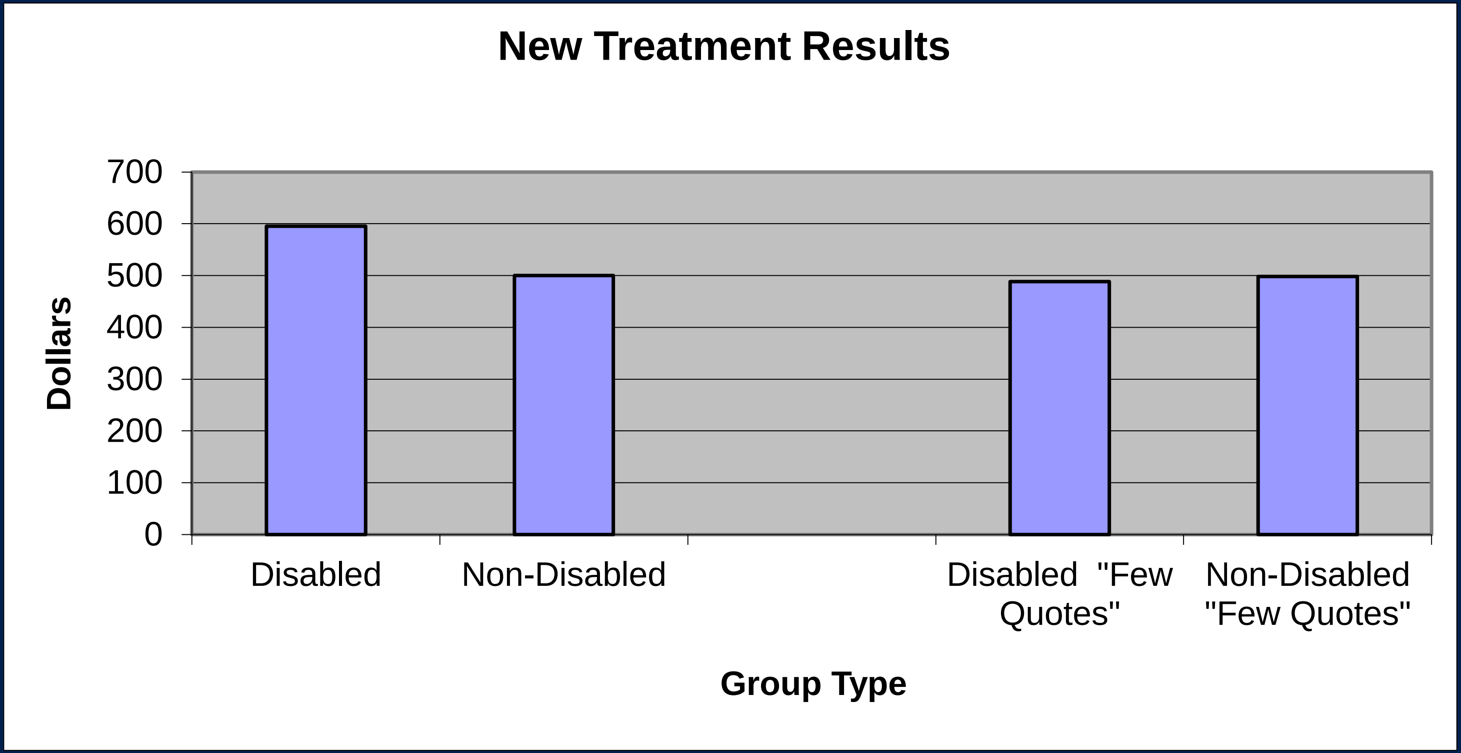
Complementary Evidence

- So what?
  - We find that price differences exist
  - But why? Is it search costs or discrimination?
- New Treatment
  - Re-send different pairs to receive price quotes
  - One treatment replicates above treatment
  - Another treatment is identical except that it has both agent types explicitly noting that “I’m getting a few price quotes today

Replication Treatment



“Few Quote” Treatment



2. Understand the Local Context

- Be an expert about the market that you are studying
  - What incentivizes people in your study/context may not be the same as what incentivizes others
- Interpreting results from an intervention is quite difficult if you don’t understand subjects’ underlying motivations

Potential Hurdles: Political

- Political difficulties
  - Politicians like to reward supporters. They have ideas about where they would like a project to go and may be reluctant to randomize
  - Individuals in the control group may be angry that they are not in the treatment group
  - NGOs and private companies may have areas they want to target and want to choose the treated group

Potential Hurdles: Ethical

- Ethical issues
  - Analogous to clinical trails--withholding the treatment from the control group
    - *When treatment demonstrated effective, make it available to the control group (worms)*
  - Institutional Review Boards
    - *Do your institutions have IRBs?*
    - *Partnering with universities, which have stringent review for all human subjects research*

3. Obtain Sufficient Sample Size

- You should have a sample size that allows you to make inference.
- Using simple power tests allow you to know what is "sufficient size" before you run your experiment.
- Fewer researchers realize that even when you reject nulls power matters.

Basic Principles of Power Calculations

- Given our regression framework
  - $Y_i = \alpha + \beta T_i + \epsilon_i$
  - The treatment effect is $\widehat{\beta}$
- The variance of $\hat{\beta}$ is
  - $\frac{1}{N_T(1-N_T)}\frac{\sigma^2}{N}$
- We want to test the hypothesis
  - $H_0\colon \hat{\beta} = 0$
- The significance level, or size, of a test represents the probability of a Type I error

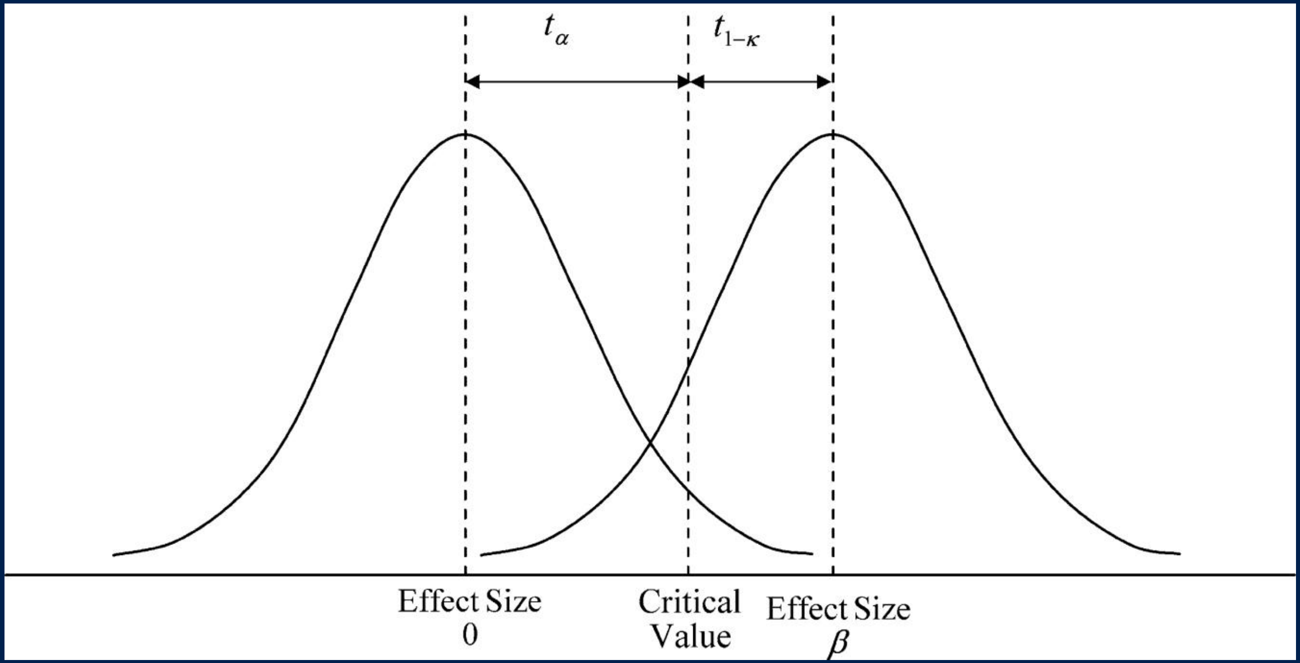
Error Types

- Type I
  - We reject the hypothesis when it is in fact true
  - False positive
- Type II
  - We fail to reject the hypothesis when it is in fact false
  - False negative

Power

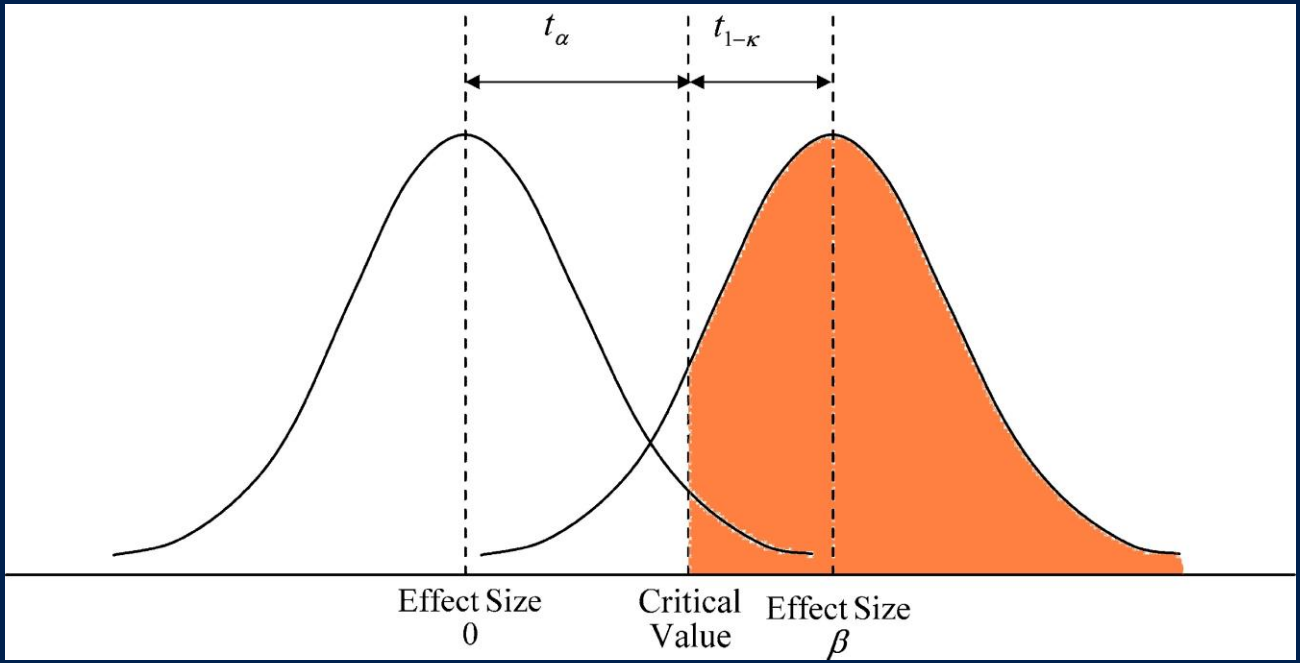
- The usual approach stems from the standard regression model: under a true null what is the probability of observing the coefficient that we observed?
- Power calculations are quite different, exploring if the alternative hypothesis is true, then what is the probability that the estimated coefficient lies outside the 95% CI defined under the null.

Hypothesis Testing



- For a given significance level H_0 will be rejected if $\hat{\beta}$ falls to the right of a critical level t_a

Hypothesis Testing



- For a given significance level H_0 will be rejected if $\hat{\beta}$ falls to the right of a critical level t_a
- The *power of the test* is the area to the right of t_a

Sample Size "Rules of Thumb"

- Assuming equal variances $\sigma_T^2 = \sigma_C^2$:

$$n_T^* = n_C^* = n^* = 2(t_{\alpha/2} + t_\beta)^2 \left(\frac{\sigma}{\delta}\right)^2$$

- Note that the necessary sample size
  - Increases rapidly with the desired significance level and power.
  - Increases proportionally with the variance of the outcomes.
  - Decreases inversely proportionally with the square of the minimum detectable effect size.

Sample Size “Rules of Thumb”

- Assuming equal variances $\sigma_T^2 = \sigma_C^2$:

$$n_T^* = n_C^* = n^* = 2(t_{\alpha/2} + t_\beta)^2 \left(\frac{\sigma}{\delta}\right)^2$$

- Sample size depends on the ratio of effect size to standard deviation. Hence, effect sizes can just as easily be expressed in standard deviations.

Sample Size "Rules of Thumb"

- Assuming equal variances $\sigma_T^2 = \sigma_C^2$:

$$n_T^* = n_C^* = n^* = 2(t_{\alpha/2} + t_\beta)^2 \left(\frac{\sigma}{\delta}\right)^2$$

- Standard is to use α=0.05 and have power of 0.80 (β=0.20).
- So to detect a one-standard deviation change using the standard approach, we would need:

$$n^* = 2(1.96 + 0.84)^2 * (1)^2 \approx 16$$

observations in each cell

Sample Size "Rules of Thumb"

- Assuming equal variances $\sigma_T^2 = \sigma_C^2$:

$$n_T^* = n_C^* = n^* = 2(t_{\alpha/2} + t_\beta)^2 \left(\frac{\sigma}{\delta}\right)^2$$

- Standard is to use α=0.05 and have power of 0.80 (β=0.20).
- So to detect a half-standard deviation change using the standard approach, we would need:

$$n^* = 2(1.96 + 0.84)^2 * (2)^2 \approx 64$$

observations in each cell

Sample Size "Rules of Thumb"

- Assuming equal variances $\sigma_T^2 = \sigma_C^2$:

$$n_T^* = n_C^* = n^* = 2(t_{\alpha/2} + t_\beta)^2 \left(\frac{\sigma}{\delta}\right)^2$$

- Standard is to use α=0.05 and have power of 0.80 (β=0.20).
- So to detect a quarter-standard deviation change using the standard approach, we would need:

$$n^* = 2(1.96 + 0.84)^2 * (4)^2 \approx 250$$

observations in each cell

Things that Effect the Power

- Grouped errors
  - Comparing between multiple groups reduces MDE
- Imperfect compliance
  - Partial compliance reduces the MDE
- Control variables
  - Controlling for baseline values increases the MDE
- Stratification
  - Blocking into similar groups increases the MDE

Power Calculations in Practice

- Many of the parameters in the power calculations are unknown
  - Need to know mean and variance in absence of experiment (get from previous lit)
  - Correlation of outcome of interest between groups (do calculations at a variety of levels).
  - The expected effect size
- Budgets are usually the binding constraint
  - Use the power calculations to help optimally design the experiment within the given budget constraint

Optimal Design

- A free, simple tool for calculating sample size
- Can do calculations and generate graphs for a number of different study designs
  - Randomized at individual level
  - Randomized at group level (clustering)
    - *With outcomes measured at individual level*
    - *Or outcomes measured at the group level*
  - Stratified or blocked designs
  - Both continuous and binary outcomes

Practical Design and Implementation Issues

Karlan, Dean. 2016. *American Economic Association* Annual Meeting

Unit of Randomization

1. Randomizing at the individual level
2. Randomizing at the group level
   "Cluster Randomized Trial"

- Which level to randomize?
  - What unit does the program target for treatment?
  - What is the unit of analysis?

How to Choose the Level

- Nature of the Treatment
  - How is the intervention administered?
  - What is the unit of intervention?
  - How wide is the potential impact?
    - *Spillovers and GE effects*
- Power requirements: larger the groups the larger the larger the total sample size
- Generally, best to randomize at the level at which the treatment is administered.

How to Choose the Level

Suppose an intervention targets health outcomes of children through info on hand-washing. What is the appropriate level of randomization?

A. Child level

B. Household level

C. Classroom level

D. School level

E. Village level

F. Don’t know

Possible Designs

- Simple lottery
- Randomization in the “bubble”
- Randomized phase-in
- Rotation
- Encouragement design

  – Note: These are not mutually exclusive.

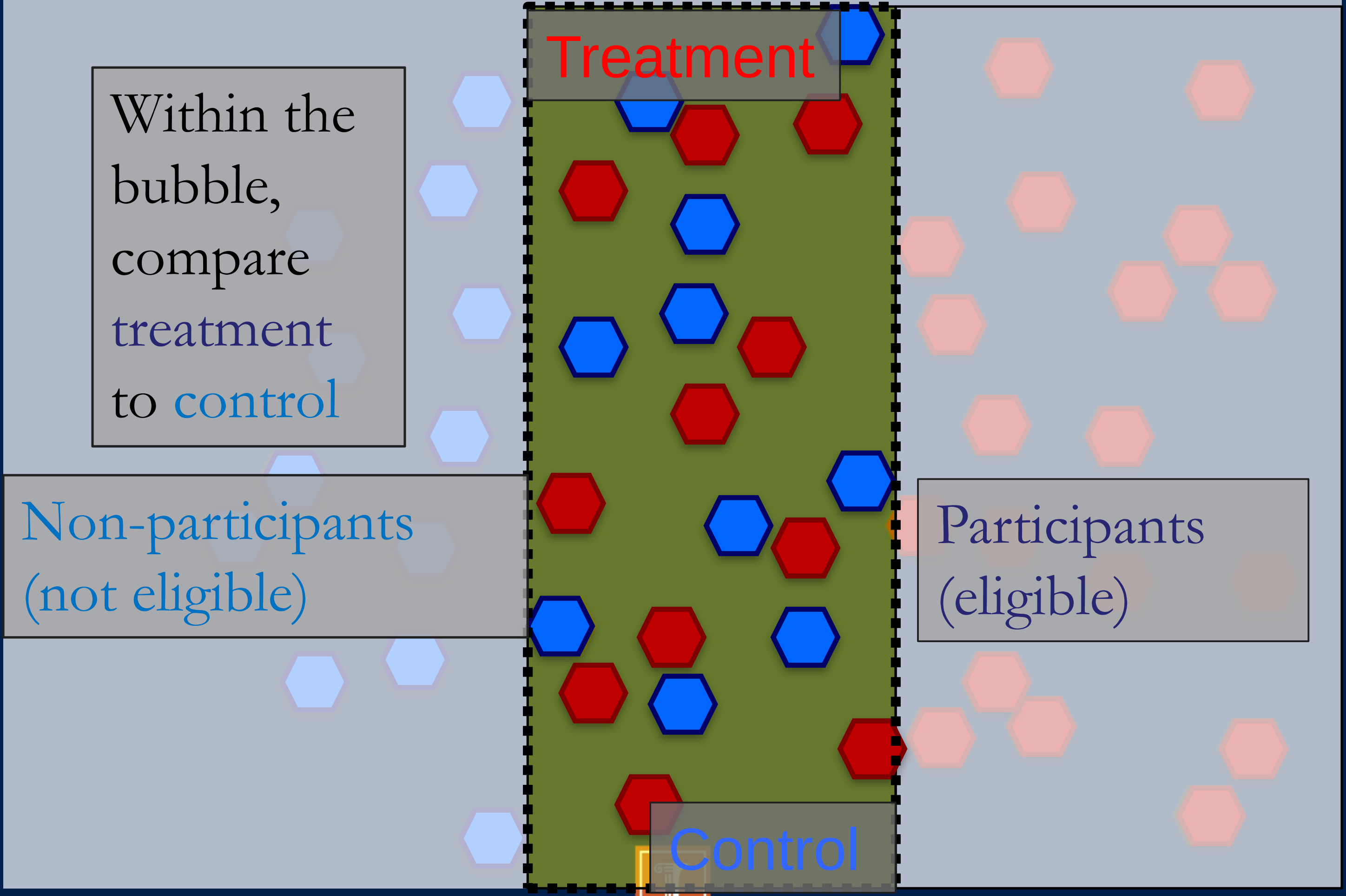
Simple Lottery

- Ideally start with a sample frame
  - Pull out of a hat/bucket
  - Use a random number generator in a spreadsheet program to order observations randomly
- With replacement?
- Proportional entry?

Randomization in “The Bubble”

- Sometimes a partner may not be willing to randomize among eligible people.
- Partner might be willing to randomize in “the bubble.”
- People “in the bubble” are people who are borderline in terms of eligibility
  - Just above the threshold → not eligible, but almost
- What treatment effect do we measure? What does it mean for external validity?

Randomization in "the bubble"



Within the bubble, compare treatment to control

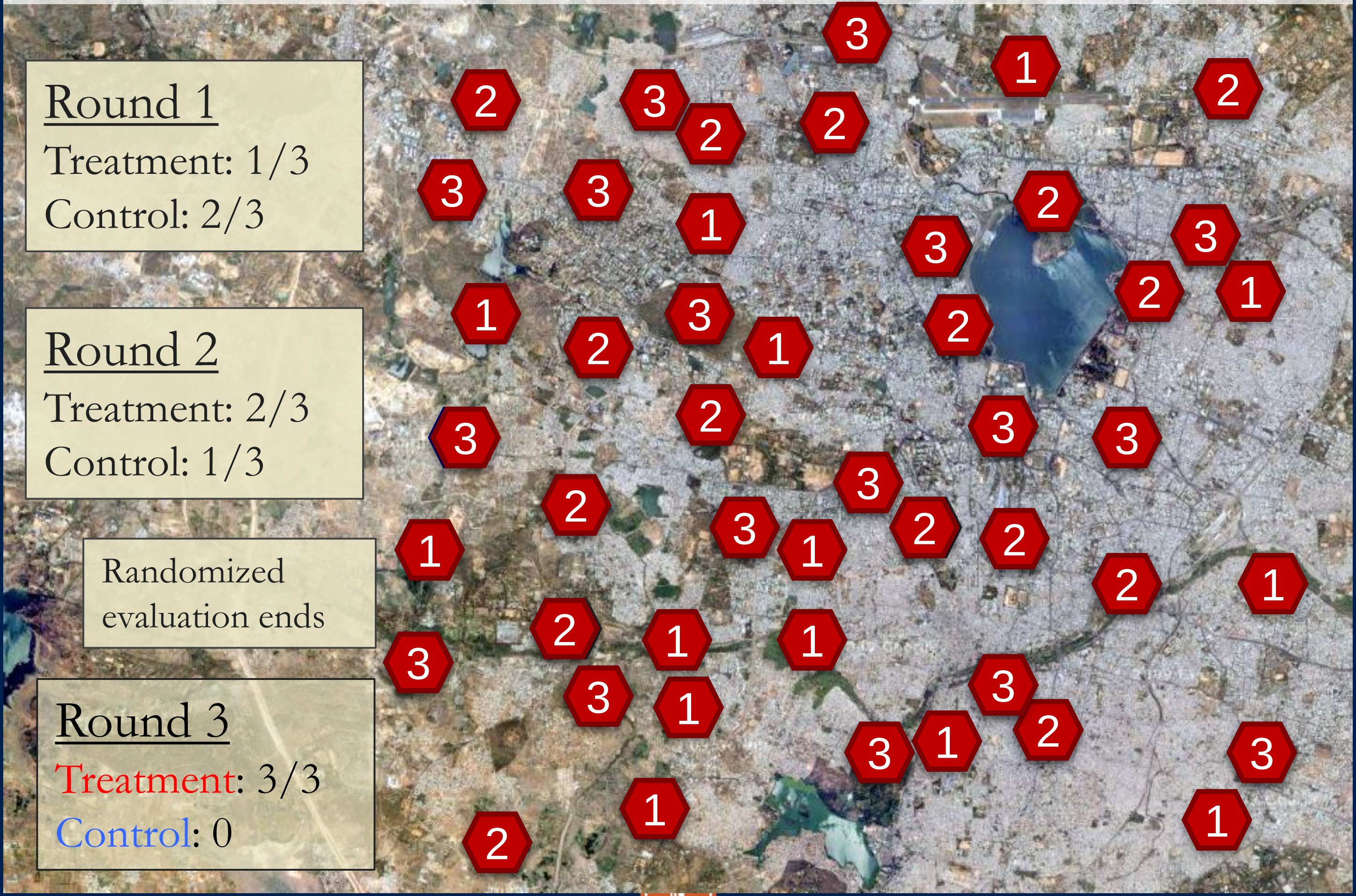
Non-participants (not eligible)

Participants (eligible)

Randomized Phase-In

- Everyone gets program eventually
  - What determines which schools, branches, etc. will be covered in which year?
- Advantages
  - Everyone gets something eventually
  - Provides incentives to maintain contact
- Concerns
  - Can complicate estimating long-run effects
  - Care required with phase-in windows
  - Do expectations change actions today?

Phase-in design



Round 1

Treatment: 1/3

Control: 2/3

Round 2

Treatment: 2/3

Control: 1/3

Randomized evaluation ends

Round 3

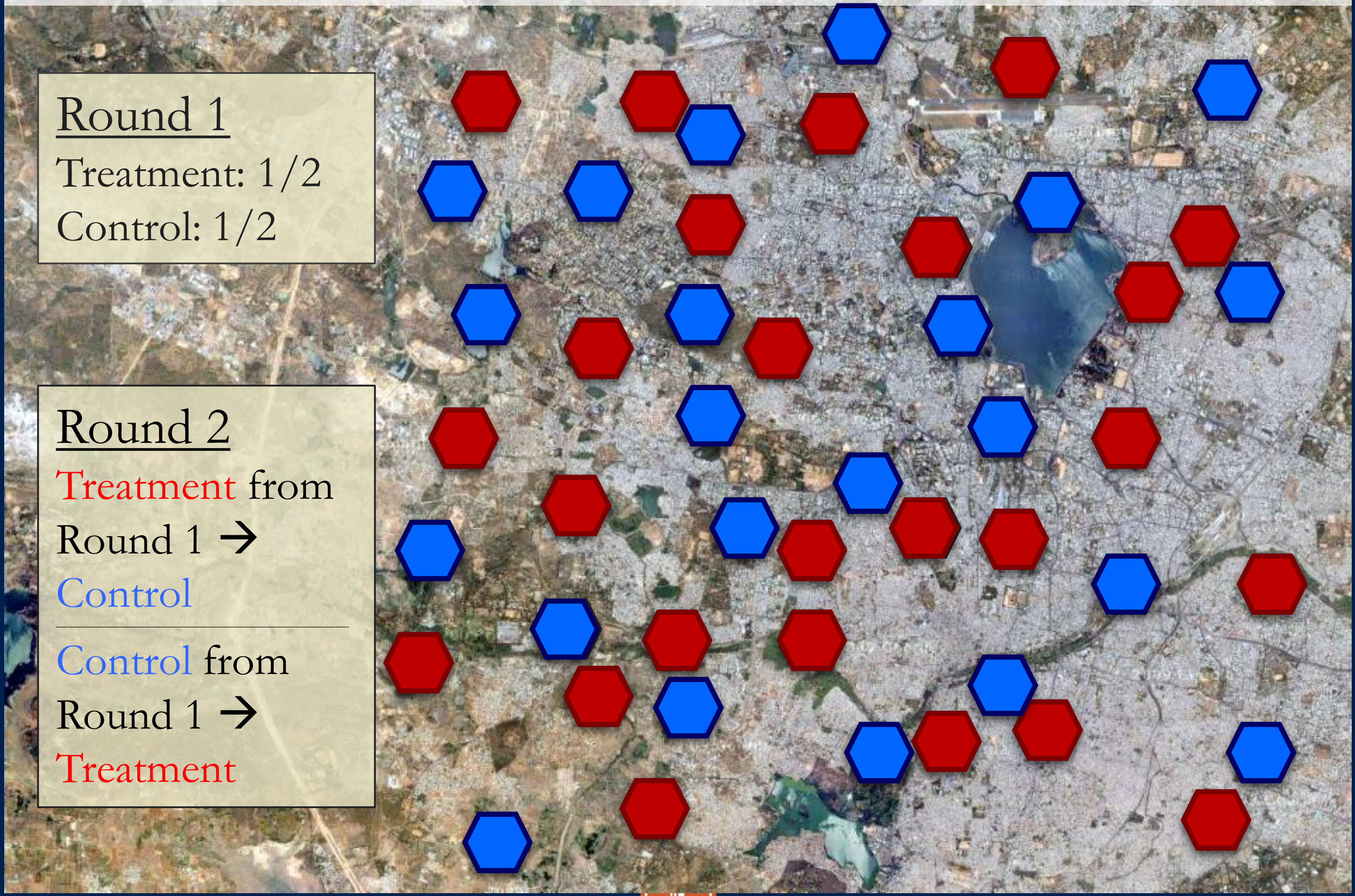
Treatment: 3/3

Control: 0

Rotation Design

- Groups get treatment in turns
- Advantages
  - Perceived as fairer; easier to get accepted
- Concerns
  - If people in Group B anticipate they'll receive the treatment the next period, they can have a different behavior in the first period
  - Impossible to measure long-term impact since no control group after first period

Rotation design



Round 1

Treatment: 1/2

Control: 1/2

Round 2

Treatment from Round 1 → Control

Control from Round 1 → Treatment

"Want to Survey Me? Then Treat Me"

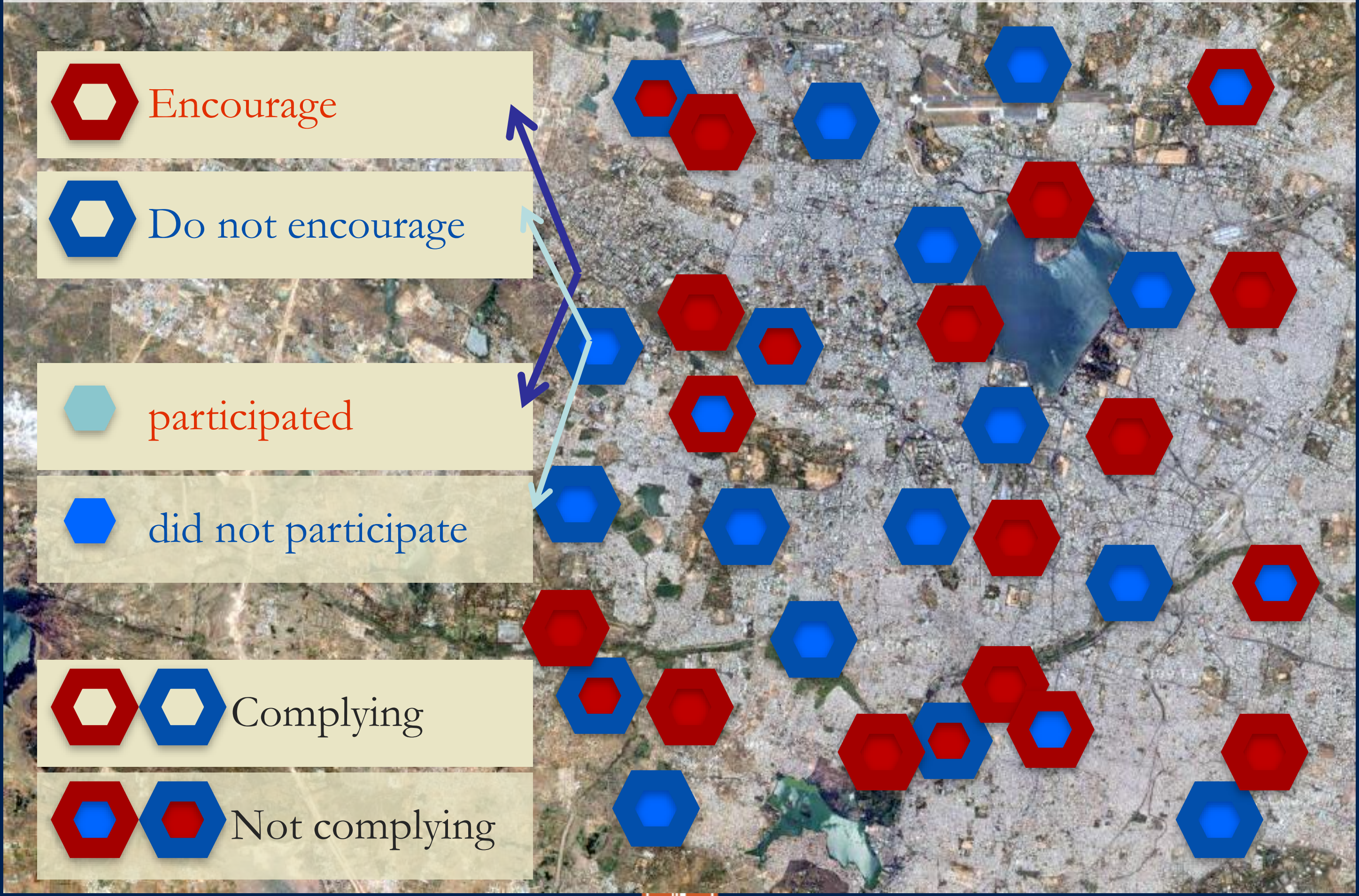
- Phase-in may not provide enough benefit to late round participants
- Cooperation from control group may be critical

- Consider within-group randomization
- All participants get some benefit
- Concern: increased likelihood of contamination

Encouragement Design

- Sometimes it's practically or ethically impossible to randomize program access
- But most programs have less than 100% take-up
- Randomize encouragement to receive treatment

Encouragement design



⬡ Encourage

⬡ Do not encourage

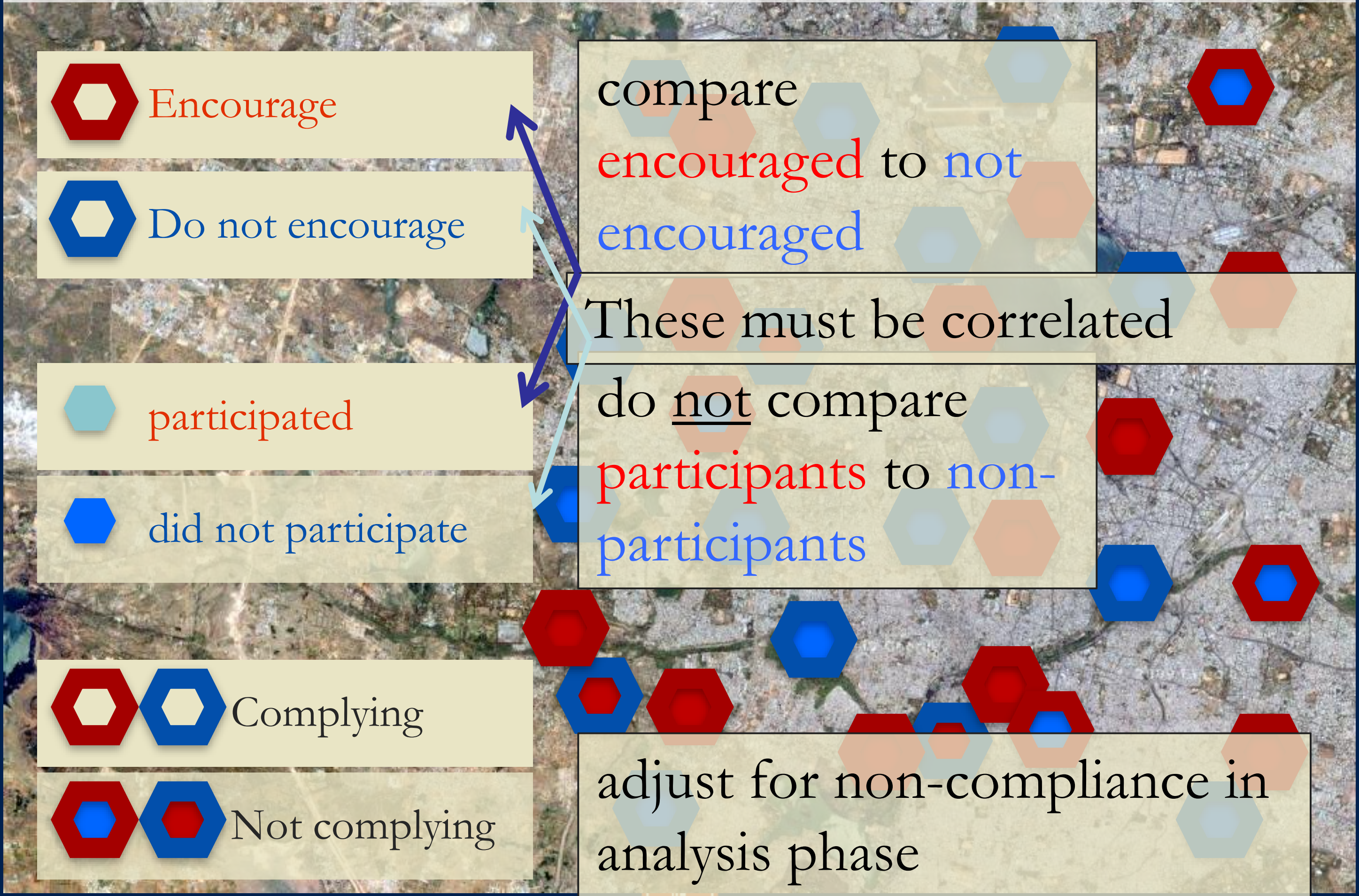
⬢ participated

⬢ did not participate

⬡⬡ Complying

⬡⬡ Not complying

Encouragement design



⬡ Encourage

⬡ Do not encourage

⬢ participated

⬢ did not participate

⬡⬡ Complying

⬡⬡ Not complying

compare encouraged to not encouraged

These must be correlated

do not compare participants to non-participants

adjust for non-compliance in analysis phase

What Is “Encouragement”?

- Something that makes some folks more likely to use program than others
- Not itself a “treatment”
- For whom are we estimating the treatment effect?
- Crucial:
  - Think about who responds to encouragement
  - Are they different from the whole population?

Stratification or Blocking

- Objective: balancing your sample when you have a small sample

- What is it:
  - Dividing the sample into different subgroups
  - Assigning treatment and control with precise proportions, within each subgroup

When to Stratify

- Stratify on variables that could have important impact on outcome variable
- Stratify on subgroups that you are particularly interested in (where may think impact of program may be different)
- Stratification more important with small sample frame
- Can get complex to stratify on too many variables
- Makes the draw less transparent the more you stratify

Varying Levels of Treatment

- Some schools are assigned full treatment
  - All kids get pills
- Some schools are assigned partial treatment
  - 50% are designated to get pills
- Testing subsidies and prices

Relative Size of Treatments

- All depends on relative weight of importance to the researcher
- 2 (similar) treatments and 1 control:
  - If you want to maximize the any treatment vs control test: 25/25/50.
  - If you want to maximize all pairwise tests equally: 33/33/33.
  - If you want to maximize the T1 vs T2 test: maybe 40/40/20.

Data Collection – The Baseline Survey

- In theory pure randomization renders baseline surveys unnecessary
- So, why is it still important to conduct them?
  - Generates control variables that reduce variance in outcome
  - Makes it possible to examine interactions between initial conditions and the impact of the program
  - Provides an opportunity to check if randomization was successful
  - Offers opportunity to test and refine data collection procedures

A Practical Example

- Your agency is implementing an irrigation program in several villages in a developing country
- They've asked you to design an RCT to measure the impact of the project.
  - How would you design the RCT?
    - *What would you measure?*
    - *What will you randomize over?*
    - *How many people will you include?*
  - What things could go wrong?