Policy Evaluation

Lecture 2: Introduction to Randomized Controlled Trials (RCTs)

Edmund Malesky, Ph.D.

Duke University

Organization

- Why Randomized Controlled Trials?
- Types of Randomized Experiments
  - ATE v. ITE V. TET
  - JTPA Example
  - Promotion/Encouragement Design

All the Rage

- In recent years, the use of randomized experiments has exploded in the social sciences, particularly in the field of applied micro-economics and developmental economics.

- Recently, randomized experiments have begun to gain currency among development practitioners and political scientists as well.

The Gold Standard

- Randomized Controlled Trials (RCT), while not universally accepted are beginning to gain currency as the "gold standard in policy evaluation."

- At the forefront of this debate has been the Jameel Poverty Action Lab (J-PAL) at MIT, led by Esther Duflo and Abhijit Banerjee.

- J-PAL commented famously only 2% of World Bank projects are properly evaluated using RCT, setting off an intense debate.

- The debate still continues….

Combines Multiple Skills

- Pulling off an RCT, involves nearly all of the tools we will add to our tool kit this semester.
  - Identifying research objectives
  - Asking the right question
  - Survey design
  - Sampling techniques
  - Case Selection
  - Econometric Analysis
    - If done right, however, you don't need high-powered tools.

Brief History

- Not really new in the social sciences
- Psychologists were performing experiments in the 1800s
- Harold Gosnell began used experiments in his work on machine politics in New York.
  - "Does canvassing increase voter turnout?"
  - Randomly assigned city blocks to receive mailed reminders
  - Turnout up 1% in pres elec of 1924, up 9% in mun elec of 1925
- But more broadly, political science/economics had delayed use.
- While political science had a 'behavioralist' revolution in the 50s and 60s. They tended to rely on surveys and not experiments.
  - Study of specific "real world" behavioral domain (unlike psychology)
  - Traditionally & reasonably worried about artificiality
  - Experimentation, by introducing artificiality is suspect
  - Control and manipulation not always possible for research

Partners in Truth-Seeking

1. **<u>Government:</u>** Intended for all population, but may want to pilot before scaling up.

    - Working with government requires high-level consensus and may face difficulties from officials who may have constituents upset about discrimination.

2. **<u>NGOs:</u>** Less subject to discrimination problems, because their programs are always isolated and individualized to some extent.

    - Are results dependent on an impossible to replicate organizational culture?

3. **<u>Multilaterals</u>** like the World Bank, Asian Development Bank...

4. **<u>For profit firms:</u>** Especially in the word of micro-credit.

Why Randomized Experiments?
Recall: Causation and Counterfactuals

- Key to successful program evaluation → estimate counterfactual by finding valid comparison groups

- Invalid comparison group → estimates of program effects mixed with estimates of other differences

- 2 methods particularly likely to give counterfeit counterfactual:
  1. Comparing outcomes of participants *before & after* program
  2. Compare outcomes of those *with & without* program

- By contrast, randomization is the gold standard of impact evaluation

- Random assignment provides robust estimate of counterfactual

Benefits of Randomization

- Random assignment among eligible provides <u>fair & transparent</u> rule

- Budget / capacity constraints, so often don't fully reach intended population.

- Ration by chance, rather than observables or first-come, first-serve.

- Random assignment yields two groups with high probability of being statistically identical, if sufficient N

- If N large, random assignment yields statistically equivalent averages for observables and unobservables

- Thus, random assignment provides valid comparison group (counterfactual).

Recall: Potential Outcomes Framework

▶ Outcome

Y_i = Observed outcome for unit i

▶ Treatment

D_i: Indicator of whether unit i received treatment

$$D_i = \begin{cases} 1 & \text{unit } i \text{ received treatment} \\ 0 & \text{unit } i \text{ did not receive treatment} \end{cases}$$

▶ Potential Outcomes

Y_{1i} : Potential outcome for i with treatment

Y_{0i} : Potential outcome for i without treatment

Random Assignment & Selection Bias

▶ Comparing outcomes for the treated and untreated often yields incorrect estimates

$$E[Y|D = 1] - E[Y|D = 0]$$

$$E[Y_1|D = 1] - E[Y_0|D = 0]$$

$$E[Y_1|D = 1] - E[Y_0|D = 0] + (E[Y_0|D = 1] - E[Y_0|D = 1])$$

which rearranges to:

$$E[Y_1|D = 1] - E[Y_0|D = 1] + E[Y_0|D = 1] - E[Y_0|D = 0]$$

which consists of the Average Treatment Effect on the Treated (ATET) and Selection Bias:

$$\underbrace{E[Y_1 - Y_0|D = 1]}_{\text{ATET}}$$

$$+ \underbrace{E[Y_0|D = 1] - E[Y_0|D = 0]}_{\text{Selection Bias}}$$

Treatment randomly assigned, so outcomes for T & C differ in expectation only through exposure to treatment. Without treatment, outcomes same in expectation. Therefore, THERE IS NO **SELECTION BIAS!!**

Internal and External Validity

Random sampling & assignment key to external & internal validity

- *External validity*: evaluation sample accurately represents population of eligible units.

- <u>Random sampling</u> of population, so evaluation sample representative of population

- *Internal validity*: valid comparison group used, so no confounding factors in estimated impact

- Random assignment to T, C → comparison group statistically equivalent to T at baseline
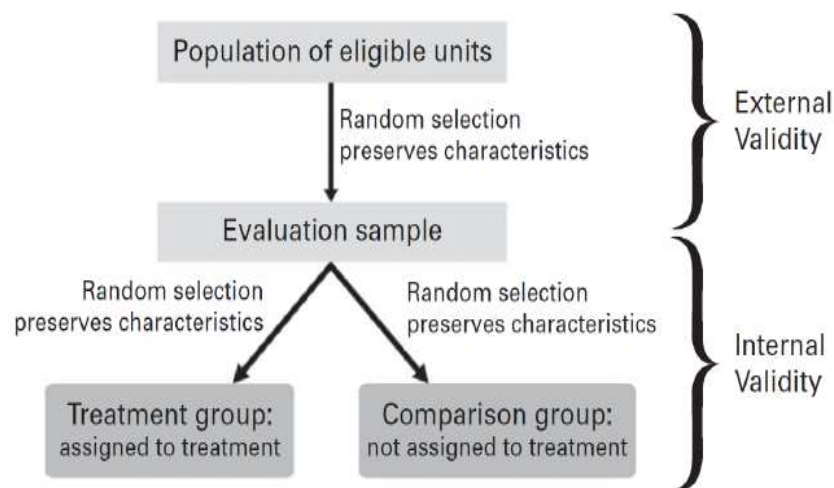
Source: Gertler, 2011.

Frontloading complexity..

So, randomization has rapidly been gaining ground in policy circles.

- If the goal of policy research is to influence policymakers, the evidence from randomized trials is very straightforward and transparent. Experiments like *Progresa* in Mexico have had huge policy effects.

- While the econometric analysis of randomized trials is completely straightforward, their use front-loads all the complexity on to the research design. Implementation is key!

- Difficult in such evaluations is not seeing what you'd like to randomize, but understanding what you will be able to successfully randomize in a given setting, and building a research design around this.

Broader Concerns about External Validity

- **External validity:** impact generalizes to other samples / populations?
  - Internal validity necessary, insufficient condition for external validity

- To generalize to population of eligible units in context, random sampling key

- Randomized evaluation fails to capture general equilibrium effects
  - Compares difference between T,C in specific area
  - Can move up to examine GE effects in village, not country/world

- Often uncertain if results extend to another country, NGO or variant

- Experiments typically conducted in small region due to logistics

- Was pilot **"gold plated,"** or can it by replicated on larger scale?
  - Difficult to learn impact for similar but not identical program
  - Should employ theory, as well as replicate, to consider generalizability

# Steps to Randomly Assign Treatment

1. **Define units eligible for program**
2. Determine sample size using power calculation
   - e.g. need larger N if minimum detectable effect small, Y rare or high variation, or if want to compare across subgroups
3. Select sample, ideally randomly
   - Use techniques from class
4. Assign T, C using transparent & ex ante rule for randomization
   - Coin, dice, lottery, random #
   - Record, or replicable w/ seed

Step 1: Eligible units

Step 2: Evaluation sample

Step 3: Randomize assignment to treatment

Comparison

Treatment

External validity

Internal validity

Ineligible   Eligible
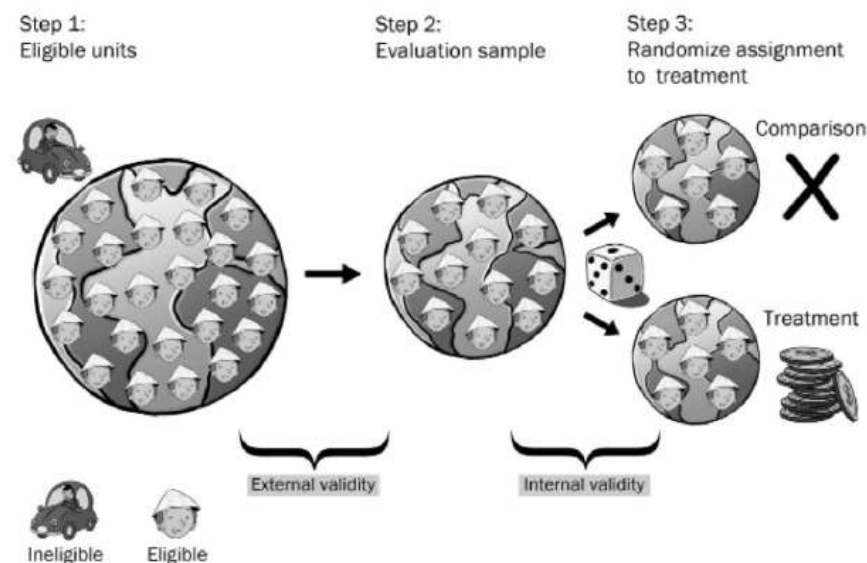
Unit if Randomization (Options)

- Two Basic Options
  - Individual Level
  - Group Level: "Cluster Randomized Trial"
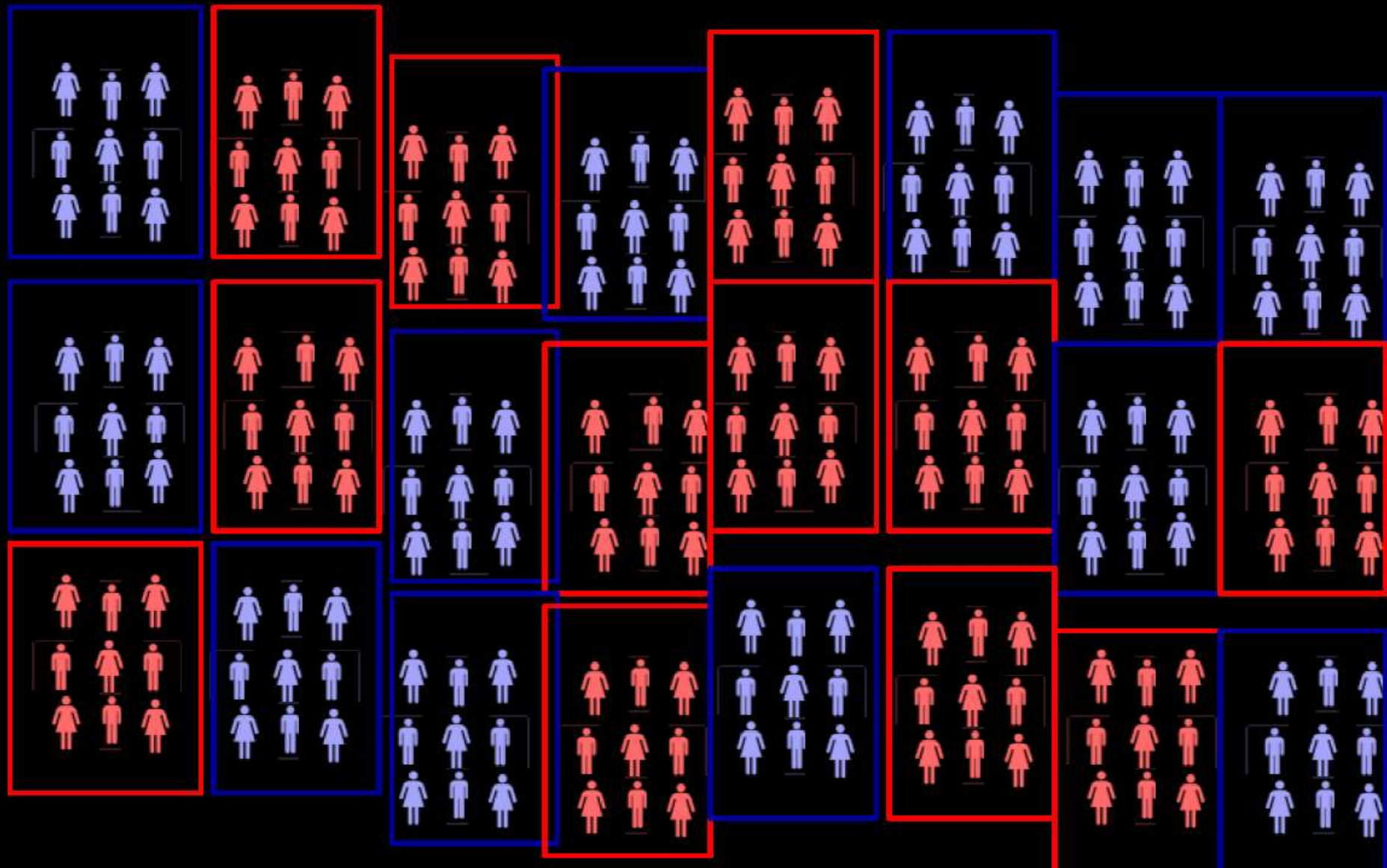
Unit of Randomization: Individual?
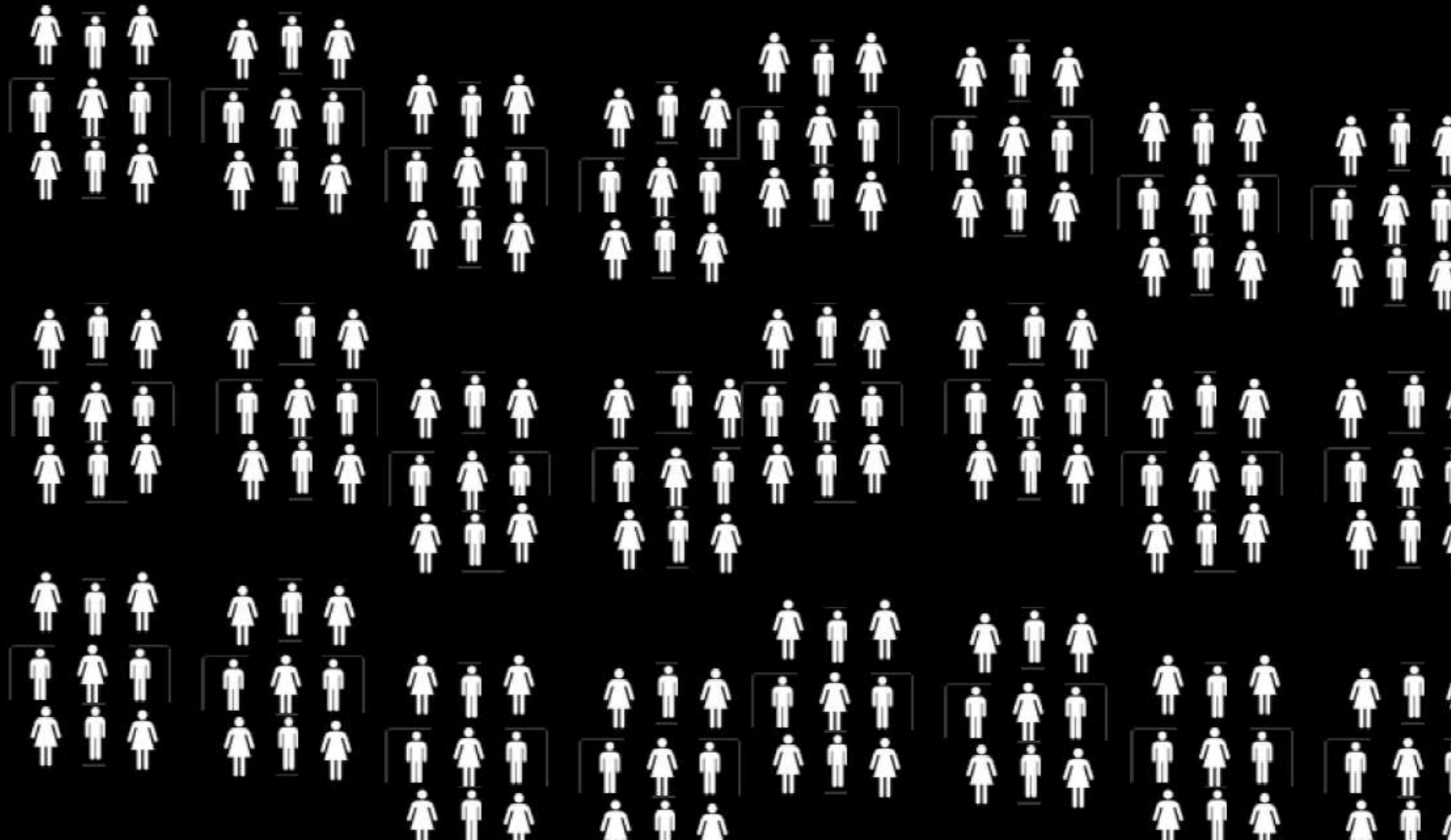
Unit of Randomization: Individual?
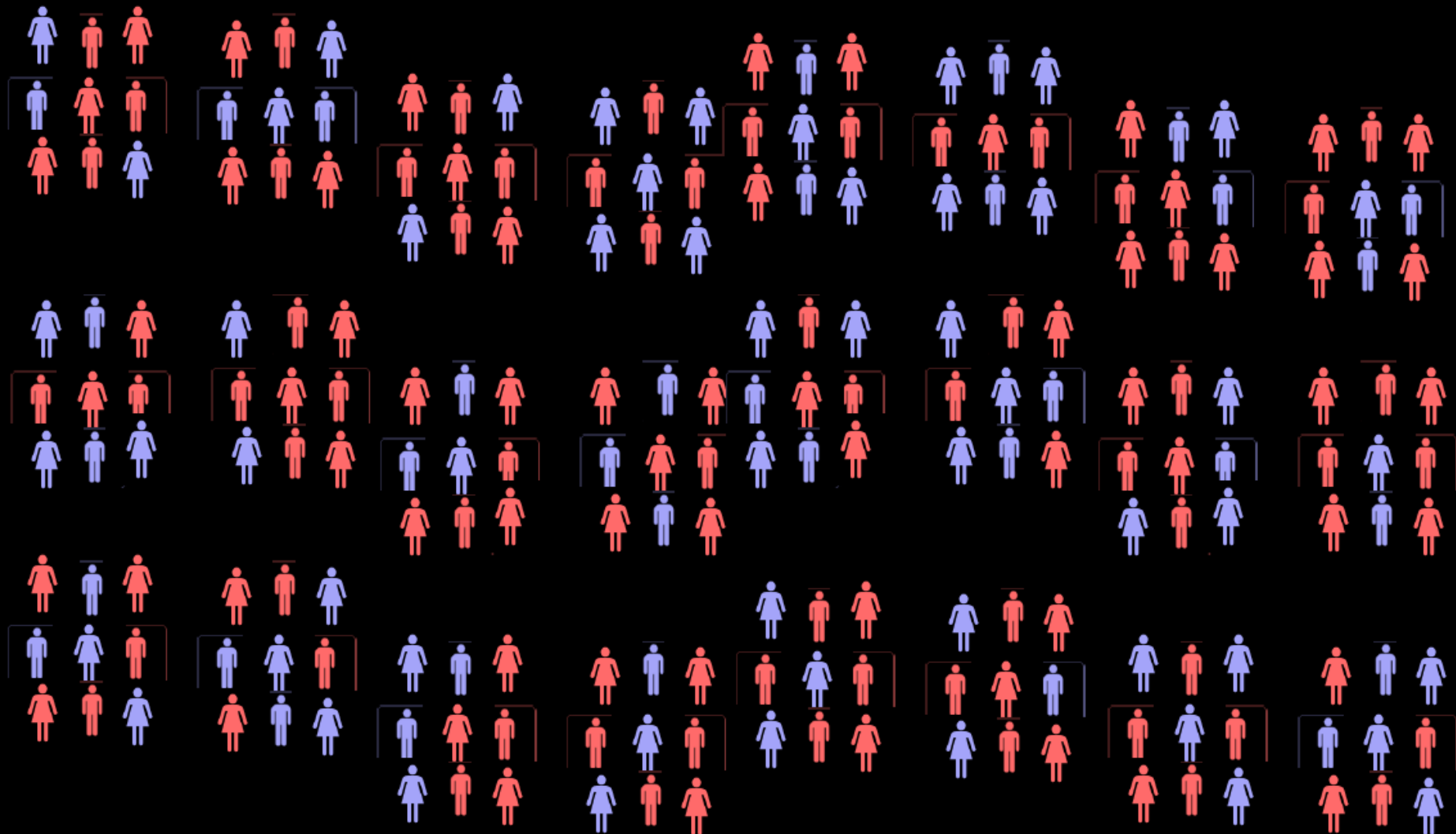
Unit of Randomization: Class?
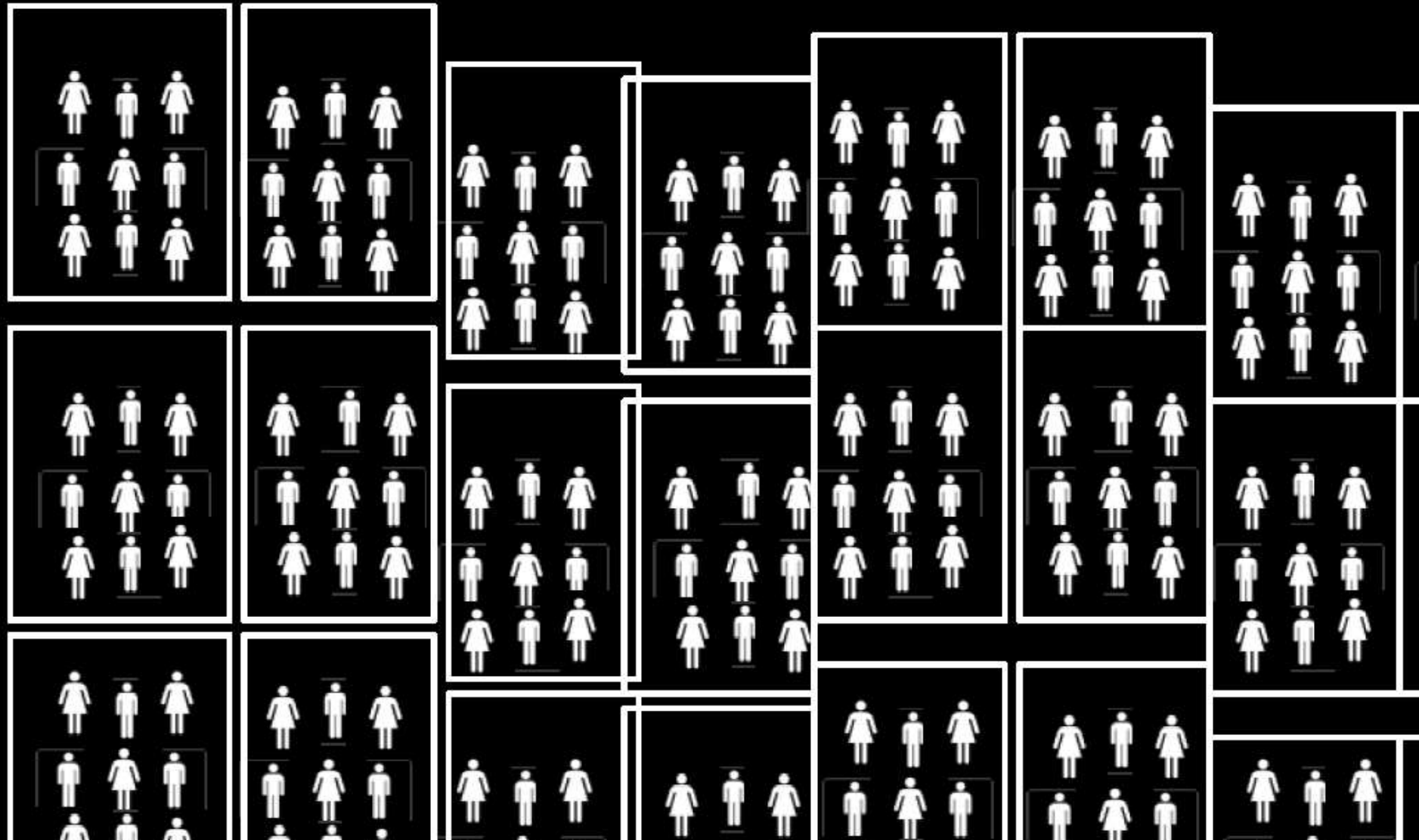
Unit of Randomization: Class?

How to Choose the Level?

- Nature of Treatment
  - How is the intervention administered?
  - How wide is the potential impact?
- Aggregation of Available Data
- Power Requirements
- Generally, best to randomize at the level at which the treatment is administered.

Methods of Randomization

1. **Classical Clinical Design**
   - Randomly allocate to treatment group(s) & control (never receive)
2. **Oversubscription Method**
   - Resources limit selection into a program. Use a lottery to determine selection into the program. Incentive of individuals is similar allowing for comparison.
3. **Randomized Order of Phase-In**
   - Program will be phased in over-time, can compare early groups to later groups (Miguel and Kremer)
     - Beware that groups not selected may change behavior out of the knowledge that the will be selected next. Both positive and negative effects.
     - Cannot study long-term effects.
4. **Within Group Randomization**
   - Provides program to different sub-group.
   - i.e. Different grades within school
   - High risk of contamination.
5. **Encouragement Design**
   - Evaluate the impact of a treatment that is available to all, but the take-up is not universal.
     - Information provision
6. **Spillover Design**
   - Take advantage of spillover, by randomling varying uptake within sample units.
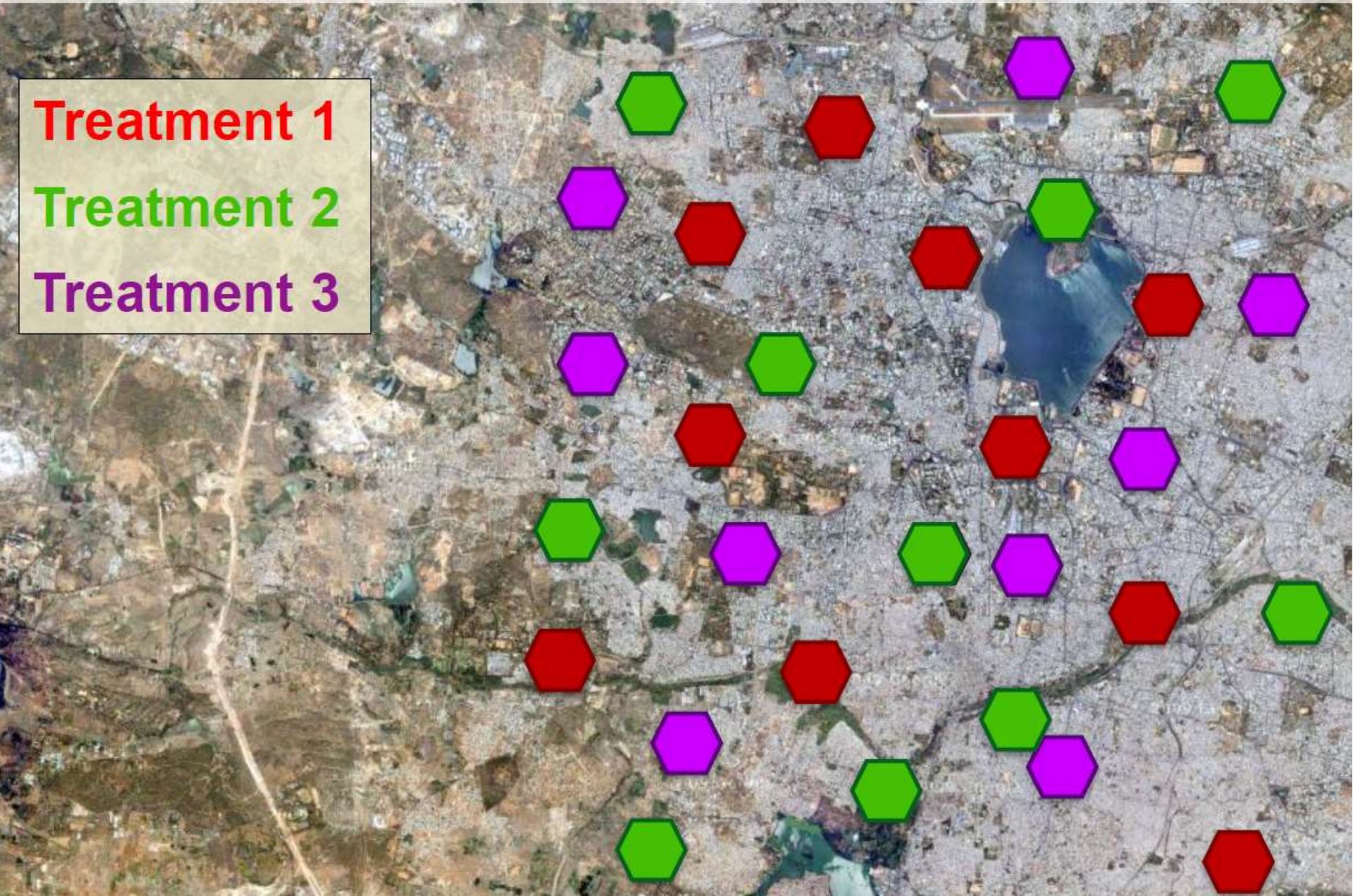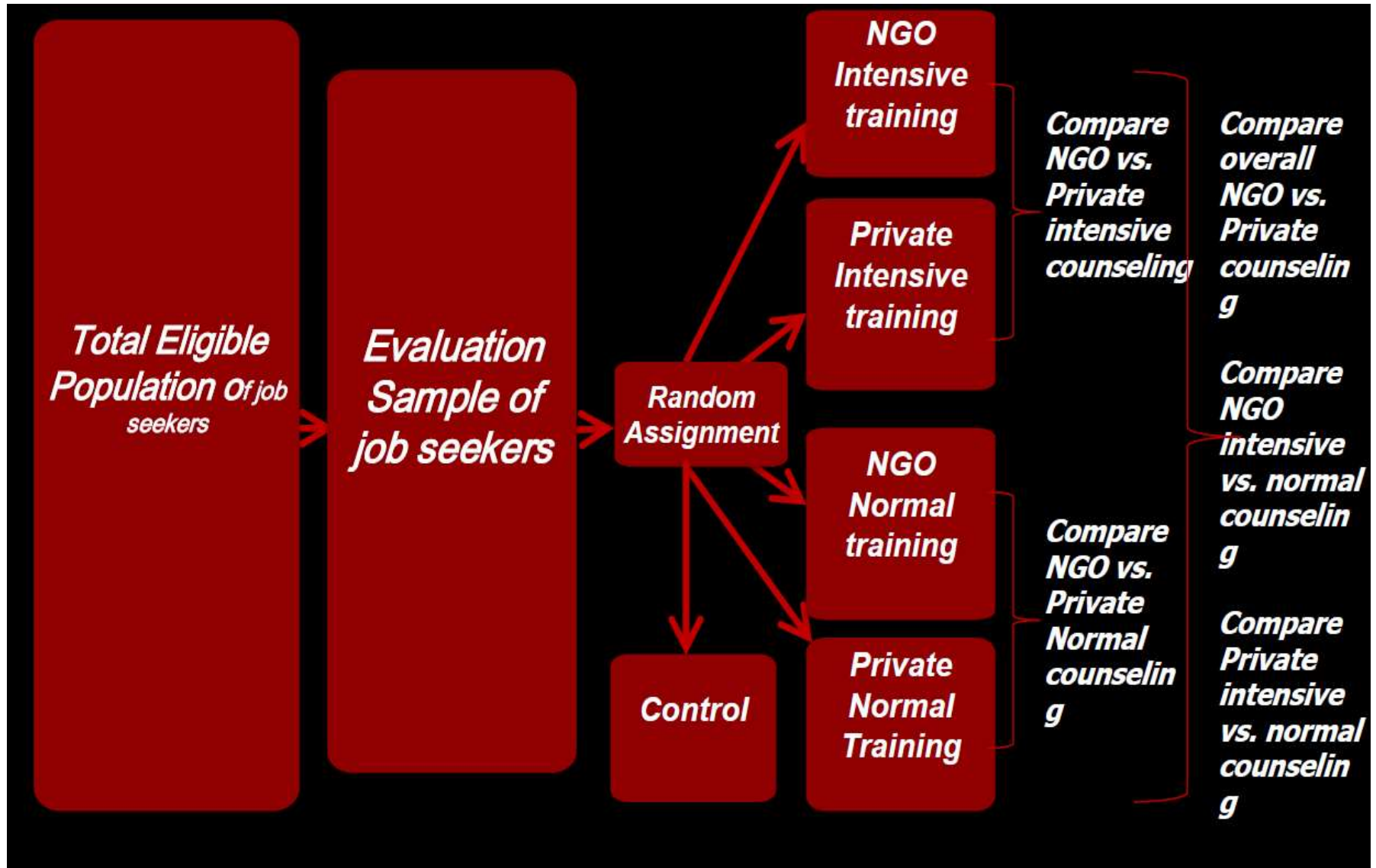   - Mali design on sugar daddies.

Multiple treatments

Treatment 1

Treatment 2

Treatment 3

Balance in Covariates

Household characteristics	Treatment villages (N = 2964)	Comparison villages (N = 2664)	Difference	t-stat
Health expenditures ($ yearly per capita)	14.48	14.57	−0.09	−0.39
Head of household's age (years)	41.6	42.3	−0.7	−1.2
Spouse's age (years)	36.8	36.8	0.0	0.38
Head of household's education (years)	2.9	2.8	0.1	2.16*
Spouse's education (years)	2.7	2.6	0.1	0.006
Head of household is female = 1	0.07	0.07	−0.0	−0.66
Indigenous = 1	0.42	0.42	0.0	0.21
Number of household members	5.7	5.7	0.0	1.21
Has bathroom = 1	0.57	0.56	0.01	1.04
Hectares of land	1.67	1.71	−0.04	−1.35
Distance to hospital (km)	109	106	3	1.02

Source: Gertler, 2011.

25

Impact: Random Assignment with Perfect Compliance

(Comparison of Means)

	Treatment	Comparison	Difference	t-stat
Household health expenditures baseline	14.48	14.57	−0.09	−0.39
Household health expenditures follow-up	7.8	17.9	−10.1**	−25.6

(Regression Analysis)

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures	−10.1** (0.39)	−10.0** (0.34)

Source: Gertler, 2011.

Non-Compliance w/Experiment

- **Most programs are voluntary, so imperfect compliance**
  - Some assigned to T don't get treatment
  - Some assigned to C obtain treatment

- **Three types of individuals can exist ("no defiers" assumption):**
  1) Compliers (enroll if T, don't enroll if C)
  2) Always-takers (enroll if T, enroll if C)
  3) Never-takers (don't enroll if T, don't enroll if C)

- **Can't identify, but share of each type in pop. flows to T,C**
  - Assignment to T increases probability of receiving T, but not by a probability of 1.
  - With certain assumptions (independence & monotonicity), Wald estimator gives effect of treatment on compliers:

$$\beta_W = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[T_i|Z_i = 1] - E[T_i|Z_i = 0]}$$
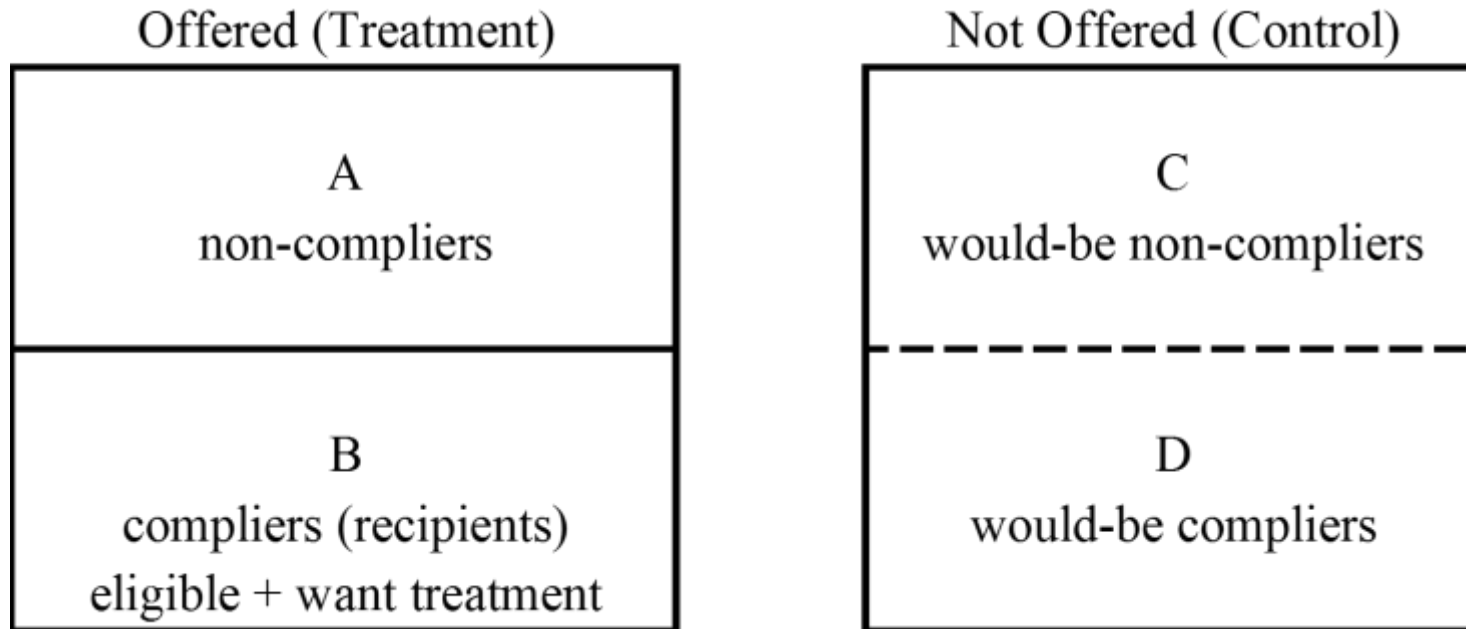
Randomized Offer

Source: Gertler, 2011

ITE and TET:

Offered (Treatment)

A non-compliers
B compliers (recipients) eligible + want treatment

Not Offered (Control)

C would-be non-compliers
D would-be compliers

- ITE compares A & B to C & D.
- Comparison of B to C & D contains selection bias even if Treatment & Control offering are randomized
- TET compares B to D, but how to establish compliance in the control?

Parsing the Idea of Random Assignment

Which do you care about, ATE, ITT or TOT?

- <u>Average Treatment Effect (ATE):</u> If the program in question is universal or mandatory:
  - What is the expected effect of treating the average individual?

- <u>Policymakers often consider intention to treat (ITT)</u> when scaling up
  - Effect takes into account that those targeted may not comply
  - High relevance, as can't usually force treatment on population
  - e.g., kids absent during deworming, too expensive to find at home
  - Helps answer: What is expected effect on an individual offered treatment, regardless whether she actually takes the treatment?

- <u>But treatment on treated (TOT)</u> estimate often important as well
  - Reveals impact of T, which you can deliver w/ other instruments
  - e.g., few take iron supplements, but want to know effect of iron
  - Corrects for fact that some T don't take it, some C take it
  - Estimate valid for compliers, not entire population
  - Helps answer: What is expected treatment effect on an individual who is offered and who takes the treatment?

Effect type determines research design:

- <u>Average Treatment Effect</u> can be directly randomized.

- <u>Intention to Treat Effect</u>:
  - Take a random sample of the entire population.
  - Conduct the normal selection process within a randomly selected subset of the sample
  - ITE given by the difference in outcomes between 'offered' and 'not offered' random samples of the population.

- <u>Treatment Effect on Treated:</u>
  - Use only the subset of the sample who would have been offered and would have taken the treatment. These are the 'compliers'.
  - Randomize treatment within the compliers.
  - TET given by the difference in outcomes between the treated and untreated compliers.
  - TET not easily estimated in practice because it requires pre-enrollment of treated units and *then* randomization. In practice, this means offering access to a lottery.

JTPA Job Training Program

- Job Training Partnership Act (JTPA): Study by Department of Labor in 1986 to examine training programs
- Applicants at 16 local JTPA programs randomly assigned to T, C
  - T allowed to enroll, C not allowed to enroll immediately
- Gathered baseline data, 2 follow-up phone surveys, state data
- Two-thirds of T actually enrolled in JTPA, <2% of C enrolled
- Not all ↑ in wages due to JTPA, as preprogram dip before applying
- Funding cut dramatically for youth programs after results published

# JTPA Job Training Program Covariate Balance (Men, Partial)

MEANS AND STANDARD DEVIATIONS

	Entire Sample	Assignment		Difference (t-stat.)
		Treatment	Control	
A. Men				
Number of observations	5,102	3,399	1,703	
Treatment				
Training	.42	.62	.01	.61
	[.49]	[.48]	[.11]	(70.34)
Outcome variable				
30 month earnings	19,147	19,520	18,404	1,116
	[19,540]	[19,912]	[18,760]	(1.96)
Baseline Characteristics				
Age	32.91	32.85	33.04	-.19
	[9.46]	[9.46]	[9.45]	(-.67)
High school or GED	.69	.69	.69	-.00
	[.45]	[.45]	[.45]	(-.12)
Married	.35	.36	.34	.02
	[.47]	[.47]	[.46]	(1.64)
Black	.25	.25	.25	.00
	[.44]	[.44]	[.44]	(.04)
Hispanic	.10	.10	.09	.01
	[.30]	[.30]	[.29]	(.70)
Worked less than 13 weeks in past year	.40	.40	.40	.00
	[.47]	[.47]	[.47]	(.56)

33

Results of Job Training (Adults)
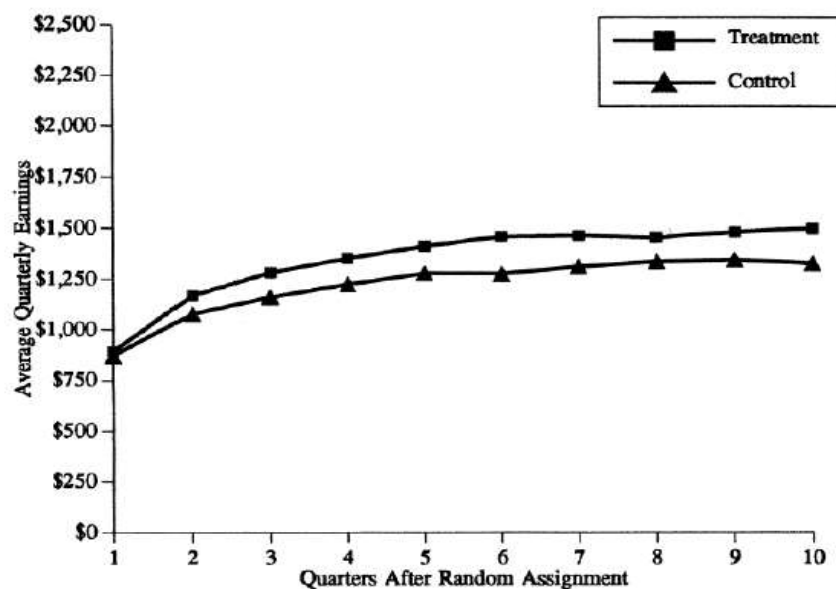
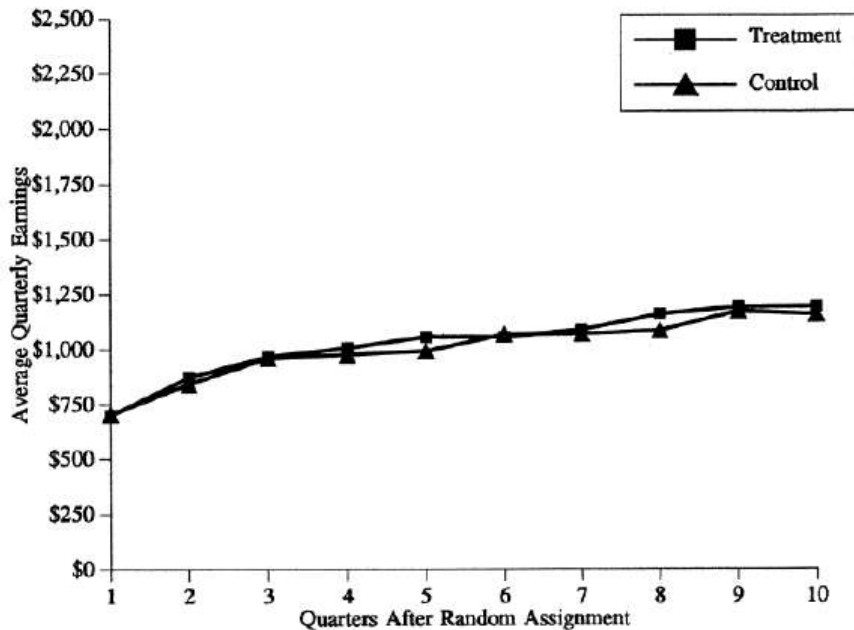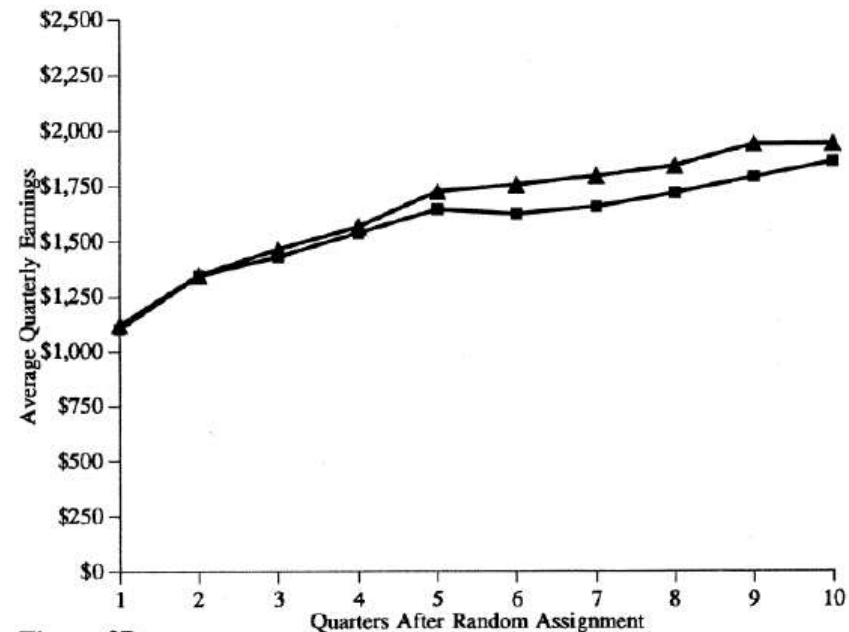

Figure 1A
Mean Earnings, by Quarter: Adult Women
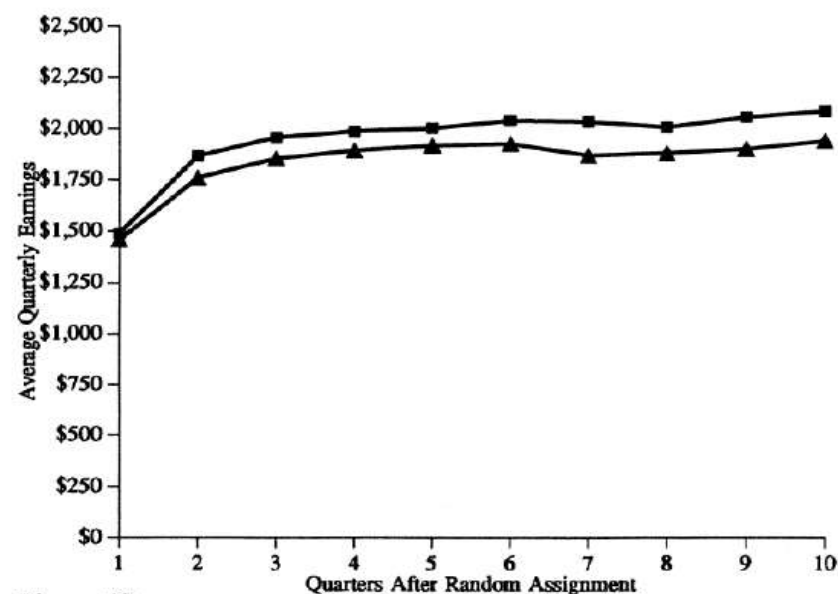
Figure 1B
Mean Earnings, by Quarter: Adult Men

Results of Job Training (Youth)

Figure 2A
Mean Earnings, by Quarter: Female Youth

Figure 2B
Mean Earnings, by Quarter: Male Youth Non-arrestees

TET v. ITE

	Enrolled in Training	Not Enrolled in Training	Total
Assigned to Training	4,804	2,683	7,487
Assigned to Control	54	3,663	3,717
Total	4,858	6,346	11,204

TET **ITE** **Corrected**

OLS AND IV ESTIMATES [OF] IMPACTS

	Comparisons by Training Status		Comparisons by Assignment Status		Instrumental Variable Estimates	
	Without Covariates (1)	With Covariates (2)	Without Covariates (3)	With Covariates (4)	Without Covariates (5)	With Covariates (6)
A. Men	3,970 (555)	3,754 (536)	1,117 (569)	970 (546)	1,825 (928)	1,593 (895)
B. Women	2,133 (345)	2,215 (334)	1,243 (359)	1,139 (341)	1,942 (560)	1,780 (532)

Note: Columns (1) and (2) show the differences in earnings by training status; columns (3) and (4) show differences by assignment status. Columns (5) and (6) report the result of using assignment status as an instrument for training. The covariates used in columns (2), (5) and (6) are *High school or GED, Black, Hispanic, Married, Worked less than 13 weeks in past year, AFDC* (for women), plus indicators for the service strategy recommended, age group and second follow-up survey. Robust standard errors in parenthesis.

Source: Abadie, 2012; Bloom et al, 1997.

Encouragement Design/Randomized Promotion

- Randomized promotion estimates impact if can't control participation
  - Budget exists, not politically or ethically OK to exclude
  - Randomly select units to offer *promo (*not treatment, as all can get)
- Info campaign or incentive increases take-up in random sample of population
  - Needs to increase enrollment, but not outcomes directly (not lots of $)
  - Promo affects *compliers*, though always-takers & never-takers exist
  - Creates equivalent of comparison group (promotion is an IV for T)
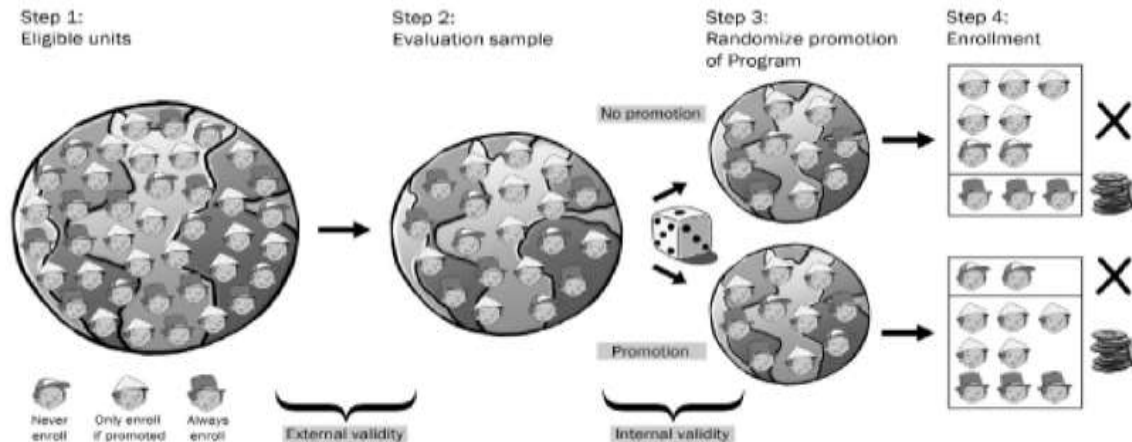
Randomized Promotion

	Promoted group	Non-promoted group	Impact
	% enrolled = 80% Average Y for promoted group = 110	% enrolled = 30% Average Y for non-promoted group = 70	Δ% enrolled = 50% ΔY = 40 Impact = 40/50% = 80
Never enroll			▬
Only enroll if promoted			
Always enroll			▬

Randomized Promotion:
Health Insurance Program

Table 4.4 Case 4—HISP Impact Using Randomized Promotion (Comparison of Means)

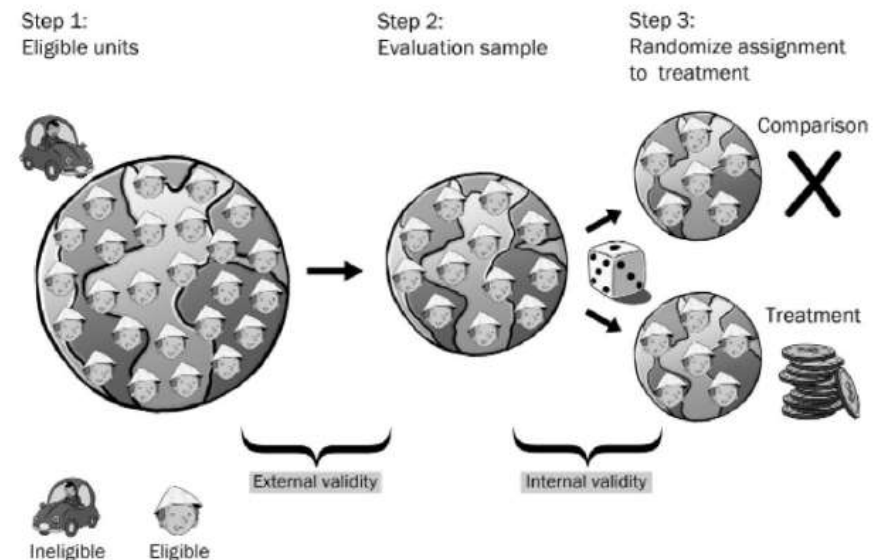|	Promoted villages	Nonpromoted villages	Difference	t-stat
Household health expenditures baseline	17.1	17.2	−0.1	−0.47
Household health expenditures follow-up	14.9	18.8	−3.9	−18.3
Enrollment in HISP	49.2%	8.4%	40.4%	

Table 4.5 Case 4—HISP Impact Using Randomized Promotion (Regression Analysis)

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures	−9.4** (0.51)	−9.7** (0.45)

Recall: Steps to Randomly Assign Treatment

1. **Define units eligible for program**
2. Determine sample size using power calculation
   - e.g. need larger N if minimum detectable effect small, Y rare or high variation, or if want to compare across subgroups
3. Select sample, ideally randomly
   - Use techniques from class
4. Assign T, C using transparent & ex ante rule for randomization
   - Coin, dice, lottery, random #
   - Record, or replicable w/ seed

Step 1:
Eligible units

Step 2:
Evaluation sample

Step 3:
Randomize assignment to treatment

Comparison

Treatment

External validity

Internal validity

Ineligible   Eligible

Conclusions

- Experiments are the ideal solution to the fundamental problem of causal inference.

- Complexity is frontloaded.

- Many different types of experiments to address problems in execution.

- Many different types of "treatment effects" that can be calculated.

- Key is to use the right tools for the job.