

# **Policy Evaluation**

## *Lecture 3: Designing and Running RCTs*

Edmund Malesky

June 29, 2020

Duke University & FUV

# Special Issues

- Measuring Impact
- Stratification and Blocking
- Power Calculations
- Checking Balance

# Basic RCT – Measuring Impact

- Data Required
  - Outcome data on treatment and control
  - Baseline data (if possible)
- Impact
  - Average Treatment Effect
    - Experiment - Counterfactual
    - Average Treatment minus Average Control

$$\begin{aligned} E[Y_1 - Y_0 | D = 1] \\ = \bar{y}_T - \bar{y}_C \end{aligned}$$

# **RANDOM ORDER OF PHASE-IN DESIGN**

# **Phase-In: Takes Advantage of Expansion**

- Ethical: Everyone gets program eventually
- Practical: Natural approach when expanding program faces resource constraints
- Randomization: What determines which schools, branches, etc. will be covered in which year?

# Features of Phase-In Design

- **Counterfactual:**
  - After year 1, people/locations starting the intervention in Yr 2, 3... serve as the control group. After year 2, the participants starting the intervention in Yr. 3, 4... serve as the control group... and so on
- **Data required:**
  - Baseline (depending) and outcome data
- **Considerations:**
  - Over time, you lose the control group
  - Possible anticipatory effects by those to receive treatment in out-years



# Phase-In Design

## Round 1

Treatment: 1/3

Control: 2/3

## Round 2

Treatment: 2/3

Control: 1/3

Randomized  
evaluation ends

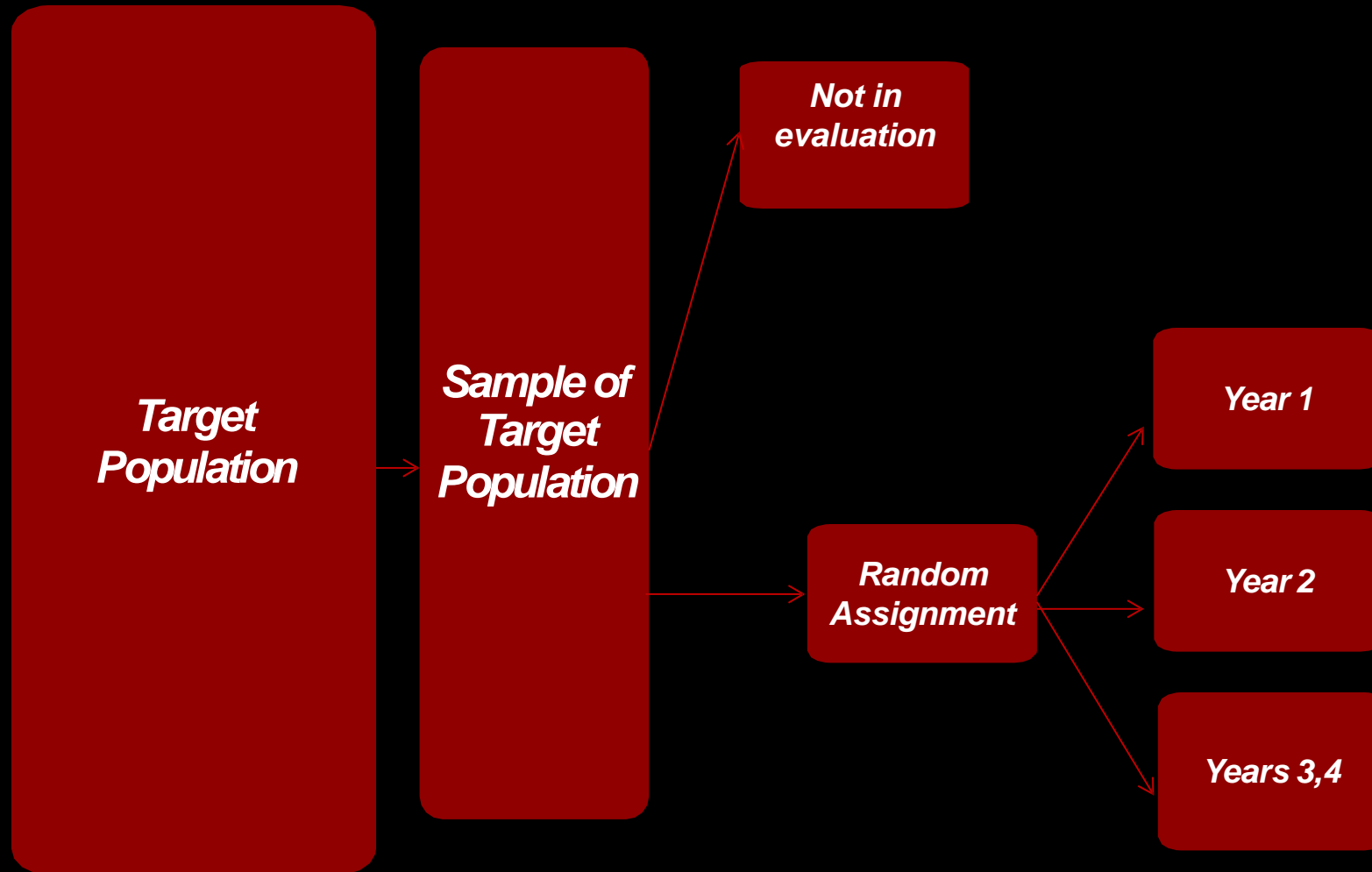
## Round 3

**Treatment:** 3/3

**Control:** 0



# RCTs | Phase In Design





# Phase-In– Measuring Impact

- Impact
  - After Year 1: Average Treatment Group (those receiving in Year 1) minus those that will receive treatment in Year 2 & 3.

$$\begin{aligned} E[Y_1 - Y_0 | D = 1] \\ = \bar{y}_{Y1} - \bar{y}_{Y2\&3} \end{aligned}$$

- After Year 2: Average Treatment Group (those receiving in Year 1 & 2) minus those that will receive treatment in Year 3 & 4.

$$\begin{aligned} E[Y_1 - Y_0 | D = 1] \\ = \bar{y}_{Y1\&2} - \bar{y}_{Y3\&4} \end{aligned}$$

# Phase-In: Pros and Cons

- Pros
  - Everyone gets something eventually Provides incentives to maintain contact
- Concerns
  - Can complicate estimating long-run effects
  - Over time, you lose the control group
  - Care required with phase-in windows
  - Do expectations of change actions today?
  - Possible anticipatory effects by those to receive treatment in out- years

# **ENCOURAGEMENT DESIGN**

# Encouragement

- *What to do when you can't randomize access?*
  - Sometimes it's practically or ethically impossible to randomize program access
  - But many programs have less than 100% take-up
  - Randomize encouragement to receive treatment

# 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?
- Think about who responds to encouragement

# RCTs | Encouragement Design

- Data required:
  - Baseline (preferably) and outcome data for encouragement and non-encouragement groups
- Considerations:
  - The encouragement has to be calibrated to substantially increase enrollment,
  - The average treatment effect may be different between those enrolled because of encouragement (what we test) and the

# Encouragement Design


 Encourage

 Do not encourage

compare  
**encouraged** to **not**  
**encouraged**

These must be correlated

 participated

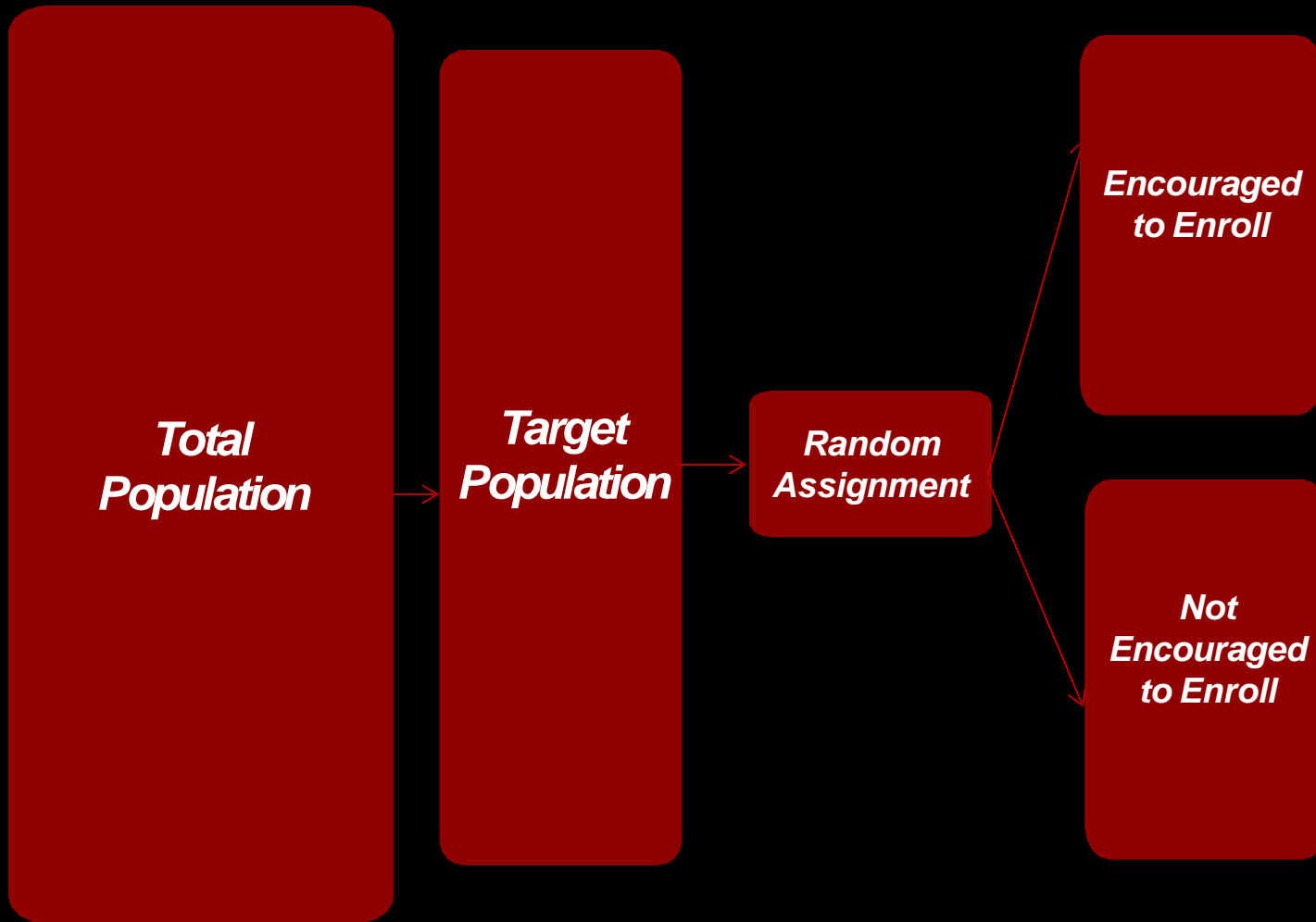
 did not participate

do not compare  
**participants** to  
**non-participants**





# RCTs | Encouragement Design



# Encouragement– Measuring Impact

- Impact

- Average Treatment Group (those with encouragement) minus those without encouragement. This is the ITE.

$$E[Y_1 - Y_0 | D = 1]$$
$$\bar{y}_E - \bar{y}_C$$

- Treatment Effect on Treated. Divide by percentage difference in enrollments (e).

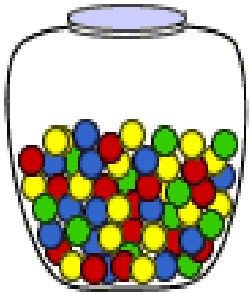
$$E[Y_1 - Y_0 | D = 1]$$
$$\frac{\bar{y}_E - \bar{y}_C}{e_E - e_C}$$

# Encouragement Design

## Impact Example

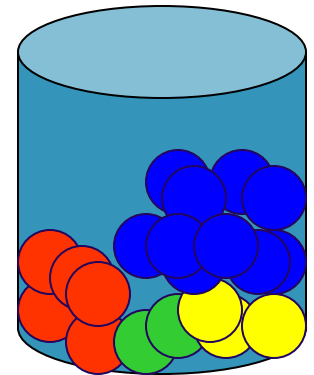
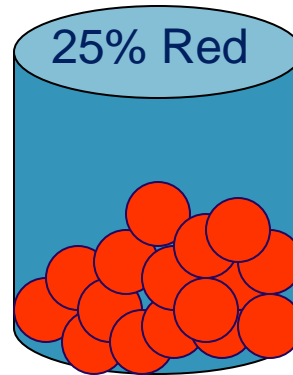
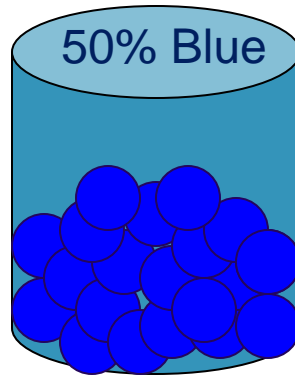
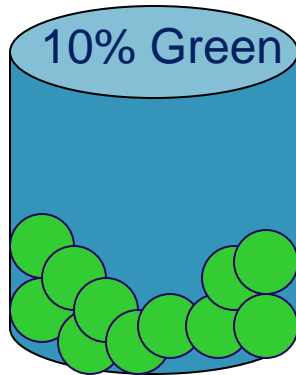
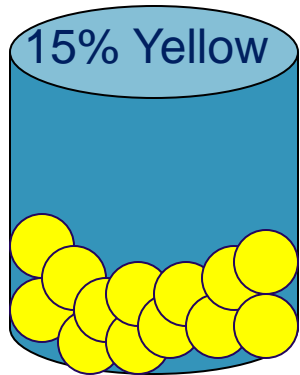
- You launch a universally available job training program and randomly assign certain areas in which individuals receive encouragement to enroll.
- You find that the overall percentage of the population that enrolls is 25% higher in encouragement areas.
- The average income in encouragement areas after one year is \$100; it is \$80 in non-encouragement areas.
- The ITE is \$20.
- The TOT of the program is therefore:  
 $(\$100 - \$80) / .25 = \$20 / .25 = \$80$

# Stratification



Simple Random Sampling: Will give me the percentage balls of a certain color (plus/minus 3%)

If I want to be more certain, I stratify and randomly sample within category



# Stratification & Blocking.

Why might you *not* want to do a single-shot randomization?

Imagine that you have a pre-observable continuous covariate  $X$  which you know to be highly correlated with outcomes.

- Why rely on chance to make the treatment orthogonal to this  $X$ ?

You can **stratify** across this  $X$  to generate an incidence of treatment which is orthogonal to this variable by construction.

What if you have a pre-observable discrete covariate which you know to be highly correlated with outcomes, or if you want to be sure to be able to analyze treatment effects within cells of this discrete covariate?

- You can **Block** on this covariate to guarantee that each subgroup has the same treatment percentage as the sample as a whole.

The expected variance of a stratified or blocked randomized estimator cannot be *higher* than the expected variance of the one-shot randomized estimator.

# When to Stratify

- When sample size is small to reduce error
  - 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 data set
- Warning 1: Can get complex to stratify on too many variables
- Warning 2: Makes the draw less transparent the more you stratify

# How to Stratify

1. Take a list of the units in the randomization frame.
2. Generate a random number for each unit in the frame.
3. Sort the list by stratification or block criterion first, then by the random number.
4. Flip a coin for whether you assign the first unit on the list to the treatment or the control.
5. Then alternate treatment status for every unit on the list; this generates  $p=.5$ .

For multiple strata or blocks:

- Sort by multiple strata or blocks and then by the random number last, then follow same as above.



# Stratified Sampling in PCI

<u>Province</u>	<u>Joint Stock Companies</u>								<u>Sole Proprietorships</u>							
	<u>Manufacturing</u>		<u>Services</u>		<u>Construction</u>		<u>Agriculture</u>		<u>Manufacturing</u>		<u>Services</u>		<u>Construction</u>		<u>Agriculture</u>	
	<u>New</u>	<u>Old</u>	<u>New</u>	<u>Old</u>	<u>New</u>	<u>Old</u>	<u>New</u>	<u>Old</u>	<u>New</u>	<u>Old</u>	<u>New</u>	<u>Old</u>	<u>New</u>	<u>Old</u>	<u>New</u>	<u>Old</u>
An Giang	46	2	18	1	58	1	25	1	236	6	115	73	750	165	34	3
Bac Can	89	0	13	1	31	0	30	0	68	3	60	0	60	1	29	0
Bac Giang	128	0	107	2	160	3	41	1	6	2	22	1	110	4	5	0
Bac Lieu	26	1	10	1	23	3	9	2	59	6	53	53	385	45	23	11
Bac Ninh	172	2	191	2	198	3	24	1	27	3	177	10	140	5	7	1
Ben Tre	19	0	18	0	31	5	8	0	142	5	109	24	706	295	181	362
Binh Dinh	44	1	46	6	156	17	37	6	134	15	120	14	503	74	74	31
Binh Duong	161	0	234	3	292	0	33	0	88	1	527	63	1227	117	27	3
Binh Phuoc	60	0	34	0	51	1	61	1	45	4	88	7	652	82	117	1
Binh Thuan	61	1	30	1	112	2	63	1	75	2	131	17	539	35	118	9
BR-VT	242	4	78	2	287	9	52	2	96	2	156	11	803	66	104	9
Ca Mau	53	0	14	0	52	2	10	0	123	8	113	48	1115	140	23	4
Can Tho	232	0	84	3	231	2	26	1	89	4	268	38	837	41	21	0
Cao Bang	28	0	9	1	25	3	29	1	127	7	12	0	56	2	17	2
Da Nang	285	5	169	10	1239	43	36	1	179	13	113	14	728	71	23	1
Dak Lak	102	2	36	1	105	6	52	0	101	7	65	3	727	88	48	1
Dak Nong	43	0	9	0	33	0	29	1	29	0	52	0	223	13	42	0

Let's look at Sample Frame for this Project

# Why wouldn't you always block or stratify?

- Bruhn & McKenzie demonstrate that the research design structure must be reflected by the treatment of standard errors in the estimation equation.
- For example, if you block on discrete values then you need to include fixed effects for these values in the estimation equation. This uses up DOF. Is this worth it?
  - Answer: it is worthwhile to block if you are blocking on a characteristic which has powerful effects on the outcome.
  - Otherwise, the blocked design eats up degrees of freedom and results in a *lower* power final test than otherwise.
- Differences between blocked or stratified & one-shot randomizations tend to disappear with a number of units greater than 300.



# STATISTICAL POWER

# Statistical Power

- Power is the probability of rejecting the null hypothesis when a specific alternative hypothesis is true.
- In a study comparing two groups, power is the chance of rejecting the null hypothesis that the two groups share a common population mean and therefore claiming that there *is* no difference between the population means of the two groups, when in fact there is a difference of a given magnitude.
- It is thus the chance of making the correct decision, that the two groups are different from each other.

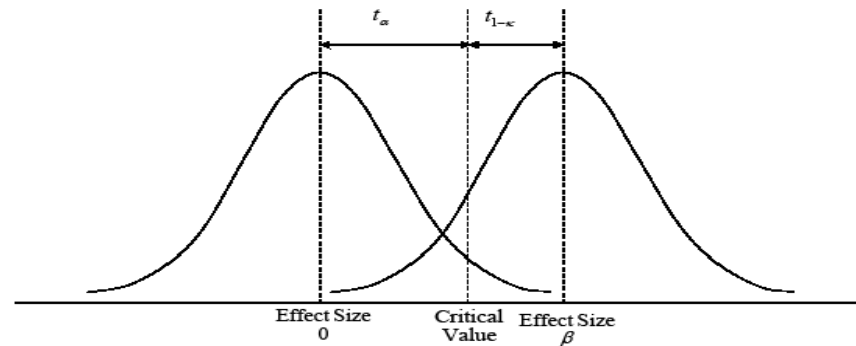
# Statistical Power

	Test result	
	“Reject Null,” Find an effect!	“Fail to Reject Null” Conclude no effect.
<b>Truth:</b> There is an effect	Great! ( $\kappa$ )	“Type II Error” (low power $1 - \kappa$ )
<b>Truth:</b> There is <b>NO</b> effect	“Type I Error” (test size $\alpha$ )	Great! ( $1 - \alpha$ )

- Probability of Type I error is “size” of test,  $\alpha$  (typically 0.05)
- Probability of Type II error is  $(1 - \kappa)$ , where  $\kappa$  is “power” of the test

# Power & Significance:

Figure 1



Left-hand curve is the distribution of beta hat under the null that it is zero,  
Right-hand curve is the distribution of beta hat if the true effect size is beta.  
Significance comes from the right tail of the left-hand curve,  
Power comes from the left-tailed distribution of the right-hand curve.

(source: Duflo & Kremer 'Toolkit')

# Power & Significance.

How many observations is 'enough'?

- Not a straightforward question to answer.
- Even the simplest power calculation requires that you know the expected treatment effect ETE,
- the variance of outcomes  $\sigma$ ,
- And the treatment uptake percentage  $p$ .
- From here, you need to pick a 'power' (the probability that you reject when you should reject, and thus avoid Type II error), typically  $\kappa = .8$  and (one-tailed).  $t_{1-\kappa} = 0.84$
- Then, pick 'significance (the probability that you falsely reject when you should accept, and thus commit Type I error), typically  $\alpha = .05$  and  $t_{1-\kappa} = 1.96$  (two-tailed).

With these you can calculate the minimum sample size as a function of the desired test power.



# Minimum Sample Size:

$$N > \frac{\sigma^2}{\left( ETE / \left( \left( t_{1-\kappa} + t_{\alpha/2} \right) \sqrt{\frac{1}{p(1-p)}} \right) \right)^2}$$

Then,

And so you can get away with a smaller sample size if you have:

- High expected treatment effects
- Low variance outcomes
- Treatment & control groups of similar sizes ( $p=.5$ )
- A willingness to accept low significance and power.

# Levels of Randomization

- Which level does one randomize at?
  - Issues:
    - The larger the groups, the larger the sample size needed to achieve a given power.
    - If spillover bias is a threat, randomization should occur at a high enough level to capture these effects.
    - Randomization at the group level can be easier to implement.
    - Individual-level randomization may create substantial resentment towards the implementing organization.

# Clustered Treatment Designs:

It is often natural to implement randomization at a unit more aggregated than the one at which data is available.

- Examples:
  - School- or village-level randomization of programs using students
  - Market- or city-level tests of political messages using voters
  - Hospital-level changes in medical practice using patients

The impact of this ‘design effect’ on the power of the test being conducted is analogous to the clustering of standard errors in the significance of regression estimators.

Bottom-line: the power of the test has more to do with the number of units *over which you randomize* than it does with the number of units *in the study*.

# Clustered Treatment Designs:

Look at the difference between the ‘minimum detectable effect’, *the smallest true impact that an experiment has a good chance of detecting*.

Without clustered design:

$$MDE > (t_{1-\kappa} + t_{\alpha/2}) \sqrt{\frac{1}{p(1-p)}} \sqrt{\frac{\sigma^2}{N}}$$

With clustered design:

$$MDE > (t_{1-\kappa} + t_{\alpha/2}) \sqrt{\frac{1}{p(1-p)J}} \sigma \sqrt{\rho + \frac{1-\rho}{n}}$$

(  $J$  is the number of equally-sized clusters,  $\rho$  is the intra-cluster correlation, and  $n$  is the obs per cluster.)

# Power calculations in practice:

- Use software!

Numerous free programs exist on the internet:

- EGAP
  - [https://egap.shinyapps.io/Power\\_Calculator/](https://egap.shinyapps.io/Power_Calculator/)
- STATA
  - power
- ‘Optimal Design’
  - [http://sitemaker.umich.edu/group-based/optimal\\_design\\_software](http://sitemaker.umich.edu/group-based/optimal_design_software)
- ‘G\*Power’
  - <http://www.pscho.uni-duesseldorf.de/aap/projects/gpower/>

Many of these use medical not social science descriptions of statistical parameters and can be confusing to use.

Make sure to use power calculator with the ability to handle clustered designs if your study units are not the same as the intervention units.

Reality check: You very frequently face a sample size constraint for logistical reasons, and then power calculations are relevant only to provide you the probability that



ToonClips.com

#1693

service@toonclips.com

# STATISTICAL BALANCE

# Post-randomization balance tests:

Quite common for researchers to write loops which conduct randomizations many times, check for balance across numerous characteristics, keep re-running until balance across a pre-specified set of statistics has been realized.

Debate exists over this practice.

Of course, generates a good-looking t-table of baseline outcomes. Functions like a multi-dimensional stratification criterion. However:

- T-tests of difference based on a single comparison of means are no longer correct, and
- It is not easy to see how to correct for the research design structure in the estimation of treatment effects (Bruhn & McKenzie, 2008).



# Example of a Balance Table

**TABLE 2. Summary Statistics**

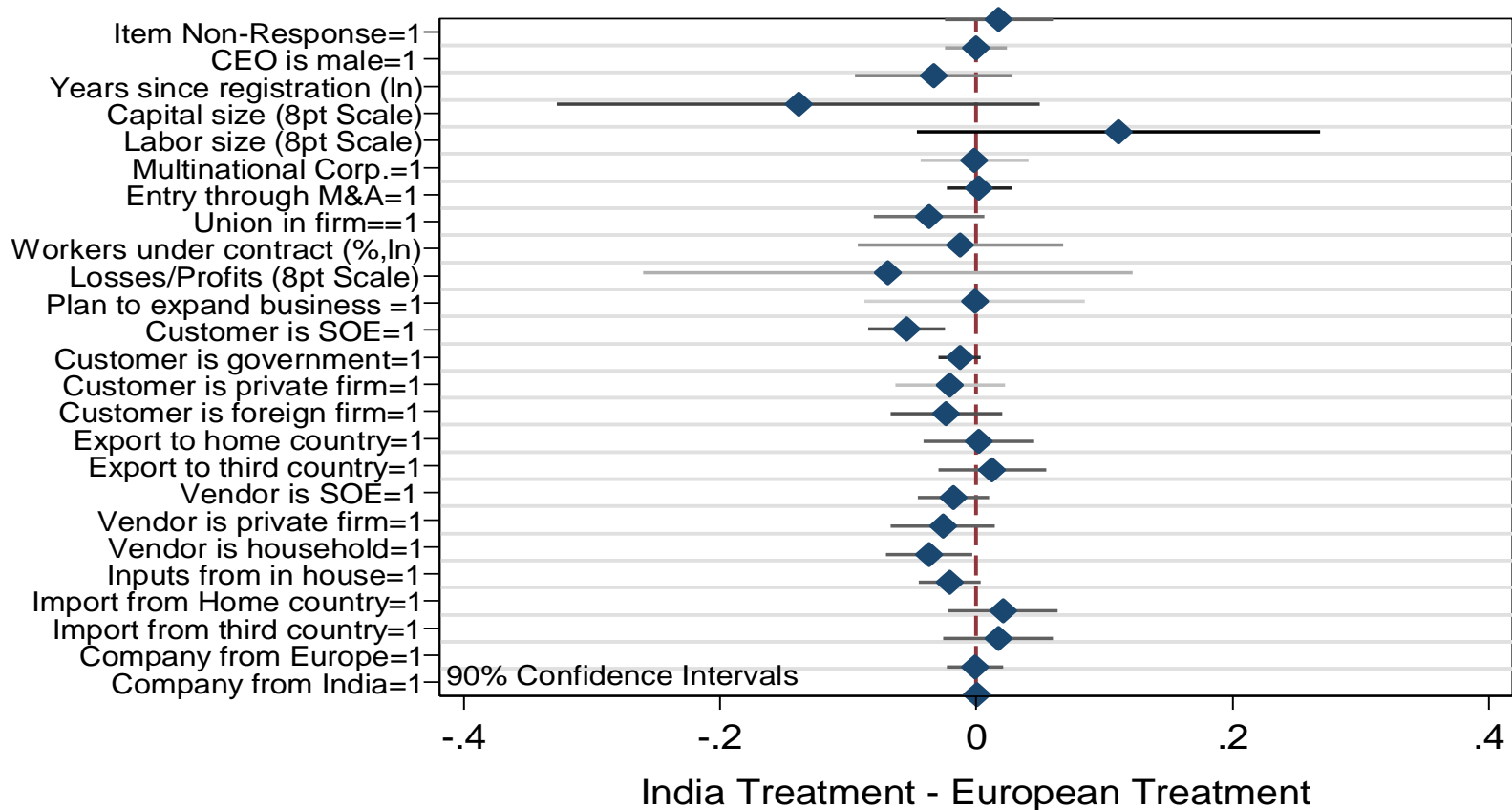
	(1)	(2)	(3)	(4)		(1)	(2)	(3)	(4)
	Mean in Meeting Group	Difference between Plebiscite and Meeting Group	p Value	Num Obs		Mean in Meeting Group	Difference between Plebiscite and Meeting Group	p Value	Num Obs
<i>Village characteristics</i>					<i>Village government characteristics</i>				
Village population (1,000 inhabitants)	2.401 [2.726]	-0.295 (0.598)	0.625	49	Village head age	45.935 [8.370]	2.368 (3.059)	0.443	47
Agricultural wage (1,000 Rupiah)	21.023 [5.892]	-1.061 (1.443)	0.466	43	Village head years of education	11.645 [2.026]	-1.409 (0.788)	0.081*	47
Percent village roads that are asphalt	0.305 [0.269]	-0.042 (0.062)	0.507	49	Number of village head candidates in last village head election	2.207 [1.013]	0.304 (0.383)	0.432	44
Number of hamlets per village	4.813 [1.839]	-0.633 (0.423)	0.142	49	More than one candidate in last village head election	0.724 [0.455]	0.089 (0.116)	0.449	44
Number of churches and mosques per village	2.438 [1.933]	-0.220 (0.563)	0.698	49	Share of population that voted in last village head election	0.888 [0.100]	-0.004 (0.031)	0.910	43
Distance to subdistrict capital (km)	5.766 [6.509]	3.548 (2.173)	0.109	49	Village head's margin of victory in last election (if challenger)	0.263 [0.262]	-0.011 (0.069)	0.870	33
Village ethnic fragmentation	0.268 [0.250]	-0.075 (0.056)	0.190	49	Number of village government executive branch members	8.516 [2.850]	-0.616 (0.703)	0.386	47
Village religious fragmentation	0.106 [0.137]	0.011 (0.051)	0.827	49	Share of hamlets represented in village executive branch	0.853 [0.240]	0.043 (0.056)	0.442	47
					Number of people in village parliament	7.750 [3.627]	-0.976 (0.832)	0.249	36
<i>Survey respondent characteristics</i>									
Survey respondent predicted log per capita expenditure	11.505 [0.279]	0.034 (0.066)	0.602	224	Share of hamlets represented in village parliament	0.843 [0.202]	0.054 (0.056)	0.339	36
Survey respondent years education	8.925 [3.088]	-0.519 (0.616)	0.404	244	Number of village parliament meetings in last year	5.714 [4.689]	-1.853 (0.878)	0.041**	44
Survey respondent is female	0.431 [0.497]	0.025 (0.023)	0.292	245	Village parliament district system (1 = district, 0 = at large)	0.241 [0.435]	0.081 (0.148)	0.587	45
Survey respondent age	41.700 [12.021]	1.896 (1.701)	0.271	245	Number of previous KDP projects	1.875 [0.976]	-0.239 (0.318)	0.455	49
Survey respondent is farmer	0.594 [0.493]	-0.052 (0.084)	0.541	245					

Notes: Column (1) presents the mean of the listed variable in the meeting villages, with standard deviations in brackets. Column (2) presents the difference between election and meeting villages, estimated with wave fixed effects, with robust standard errors in parentheses clustered at the village level. Column (3) shows the p value from a test of the null hypothesis that the listed variable is not different between elections and meeting villages. Column (4) shows the number of observations of the listed variable.

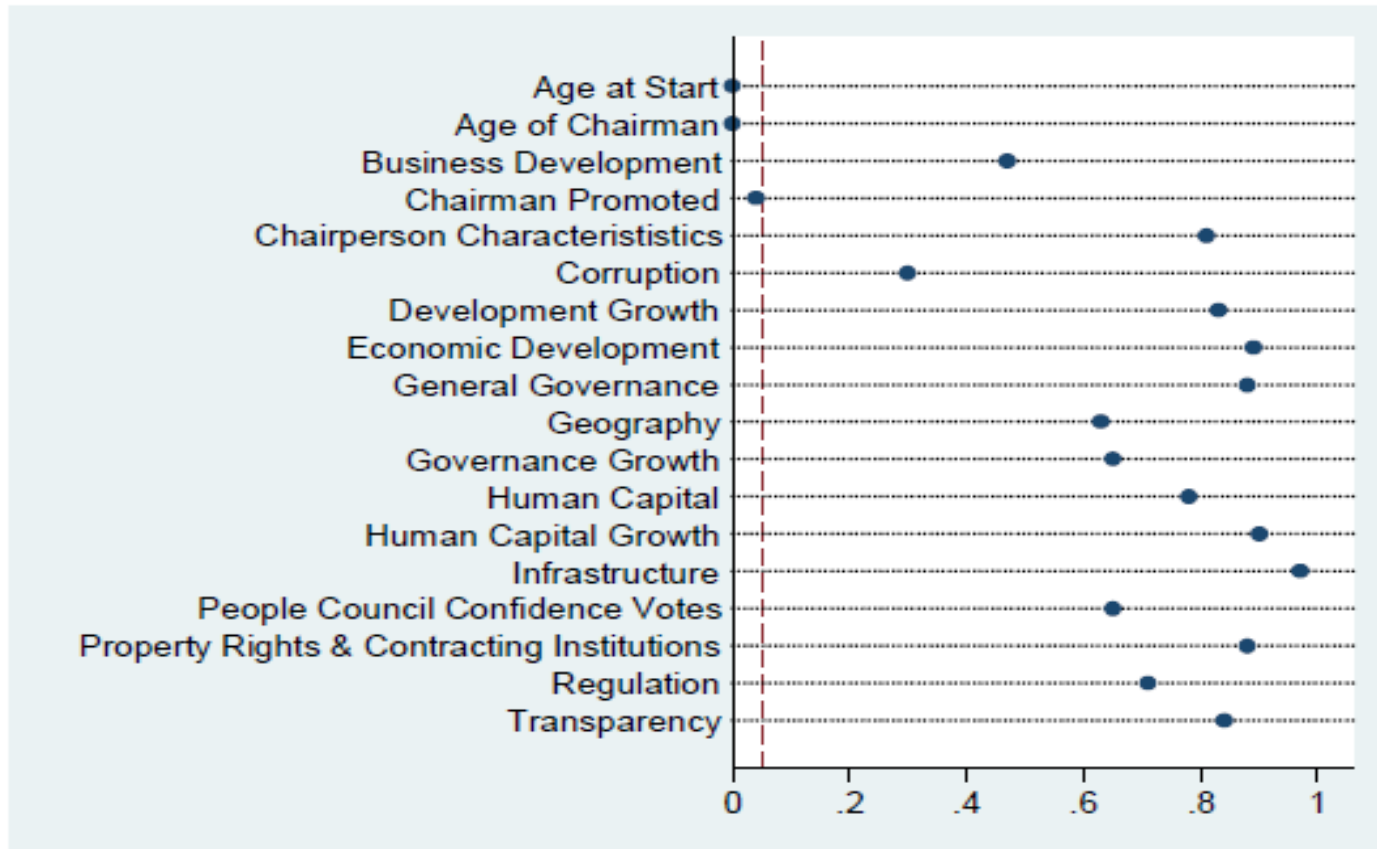
\* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%.

# Balance with CIs

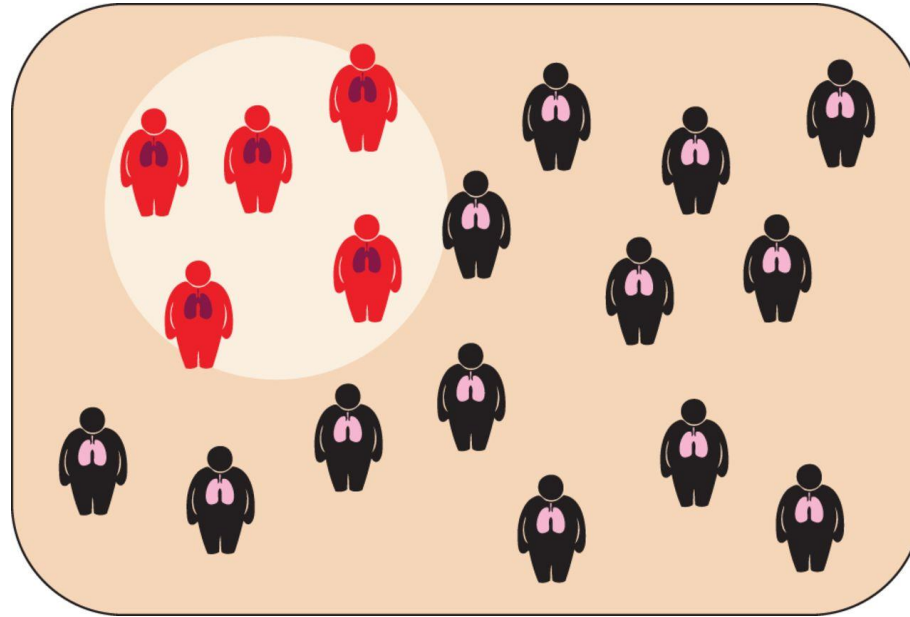
Figure 2: Survey Attrition & Balance of Confounders



# Balance with p-values



Note: Blue dots represent p-values from MANOVA analyses of grouped variables. The y-axis supplies the title of each grouping. A full list of indicators under each title can be found in Online Appendix B. Dashed line represents  $p=.05$  from the MANOVA analysis. For dots below that number, we reject the null hypothesis that the treatment and control are different on that set of criteria.



# **SUB-GROUP ANALYSIS/ HETEROGENOUS TREATMENTS**

# What is a Heterogeneous Effect?

- Any given treatment might affect different experimental subjects in different ways.
  - For whom are there big effects?
  - For whom are there small effects?
  - For whom does treatment generate beneficial or adverse effects?
- Research on such questions can help inform theories about the conditions under which treatments are especially effective or ineffective.
- It can also help inform ways of designing and deploying policies so as to maximize their effectiveness

# Conditional Average Treatment Effects (CATEs)

- A CATE is an average treatment effect specific to a subgroup of subjects, where the subgroup is defined by subjects' attributes (e.g., the ATE among female subjects) or attributes of the context in which the experiment occurs (e.g., the ATE among subjects at a specific site in a multi-site field experiment)

# Using Interaction Effects

- In addition to CATEs, researchers are also interested in treatment-by-covariate interaction effects, or the difference between two CATEs when the covariate partitioning subjects into subgroups is not experimentally manipulated.
- The coefficient  $\delta$  is the interaction effect and is interpreted as the difference between the ATE of treatment (X) among subjects in Z and the ATE of the job training program among subjects not in (Z)
- *If Z is not randomly assigned, not causal, but descriptive.*

$$Y_i = \alpha + \beta Z_i + \gamma X_i + \delta Z_i X_i + \varepsilon_i$$



# **FINAL PRACTICAL ISSUES**



# What is easily randomized?

## 1. Information:

- Trainings
- Political Message dissemination
- Dissemination of information about politician quality, corruption
- Mailers offering product variation
- Promotion of the treatment.
  - Problem with all of these is that they may be peripheral to the key variation that you really care about.
  - This has led to a great deal of research which studies that which can be randomized, rather than that which we are interested in.

## 2. Decentralized, Individual-level treatments:

- Makes evaluation of many of the central questions in policy difficult.
- Voting systems, national policies, representative-level effects, international agreements not easily tractable.
- Voter outreach, message framing, redistricting, audits much more straightforward.

# Practical Issues in the Design of Field Trials:

1. Do you control the implementation directly?
  - If so, you can be more ambitious in research design.
  - If not, you need to be brutally realistic about the strategic interests of the agency which will be doing the actual implementation. Keep it simple.
    - Has the implementing agency placed any field staff on the ground whose primary responsibility is to guard the sanctity of the research design? If not, you MUST do this.
2. Does the program have a complex selection process?
  - If so, you must build the evaluation around this process.
    - Either pre-select and estimate TET, or go for ITE.
    - If uptake is low, you need to pre-select a sample with high uptake rates in order to detect the ITE
3. Is there a natural constraint to the implementation of the program?
  - If so, use it to identify:
    - ‘Oversubscription’ method
    - If rollout will be staggered anyway, then you can often motivate the rationale behind a randomized order to the implementer.

# Imperfect Randomization

- Local Average Treatment Effect (LATE)
- Partial Compliance
  - Try to choose the design with the highest level of compliance
- Externalities
  - Spillover effects both within and outside groups.
  - If spillover is likely, design the program to address them (Miguel and Kremer).
- Attrition
  - Random attrition will impact calculation of standard errors.
  - Systematic attrition will actually bias results.
  - Make sure to record and track attrition.