RCTs - risks and considerations

Mini course on Causal inference: Lecture 4

Dr. Miri Stryjan

Department of Economics
Aalto University
2021-22

Randomized experiments - risks and considerations

- Randomizing elements of or access to a program eliminates **Selection problems** and limits **Omitted variable bias**, because the only thing that affects "treatment status" is the randomization.
- However:
  - the treatment and the control groups could still be different from each other in some ways just by bad luck.
    these differences could happen to also affect treatment response (the "effect").
  - if the sample is not very large, the researcher may also not have the statistical power to detect an effect of the treatment (the study is "underpowered").

We will talk about these points and also touch on some other important considerations related to data and measurement that you should keep in mind when reading RCT papers.

Plan for lecture

- causal comparisons and compliance.
- Balance tests and how to read balance tables.
- Statistical Power in practice.
- Additional design considerations.

Example reference

Examples in this lecture will be taken from a specific paper/project:

- Banerjee, A., Duflo, E., Glennerster, R. and Kinnan, C., 2015. The miracle of microfinance? Evidence from a randomized evaluation. American Economic Journal: Applied Economics, 7(1), pp.22-53.
- Microfinance: loans for poor people. So project essentially measures the effect of access to loans on various outcomes such as business startup, profits and household consumption.
- Big "hype" around microfinance in 2005, researchers and microfinance institution expected high take-up and large effects.

Example reference

Banerjee et al. (2015), design:

- In the project, the researchers collaborated with a Microfinance lender as they expanded into a new city: Hyderabad in 2005.
- The lender identified 104 relevant, poor neighbourhoods. Researcher randomly assigned:
  - 52 to receive a Microfinance branch (Treatment)
  - 52 remaining to serve as Control group (no Microfinance).

Causal comparisons

In the following examples, suppose we are trying to estimate the causal effect of a program or policy by comparing a treated group and a control group.

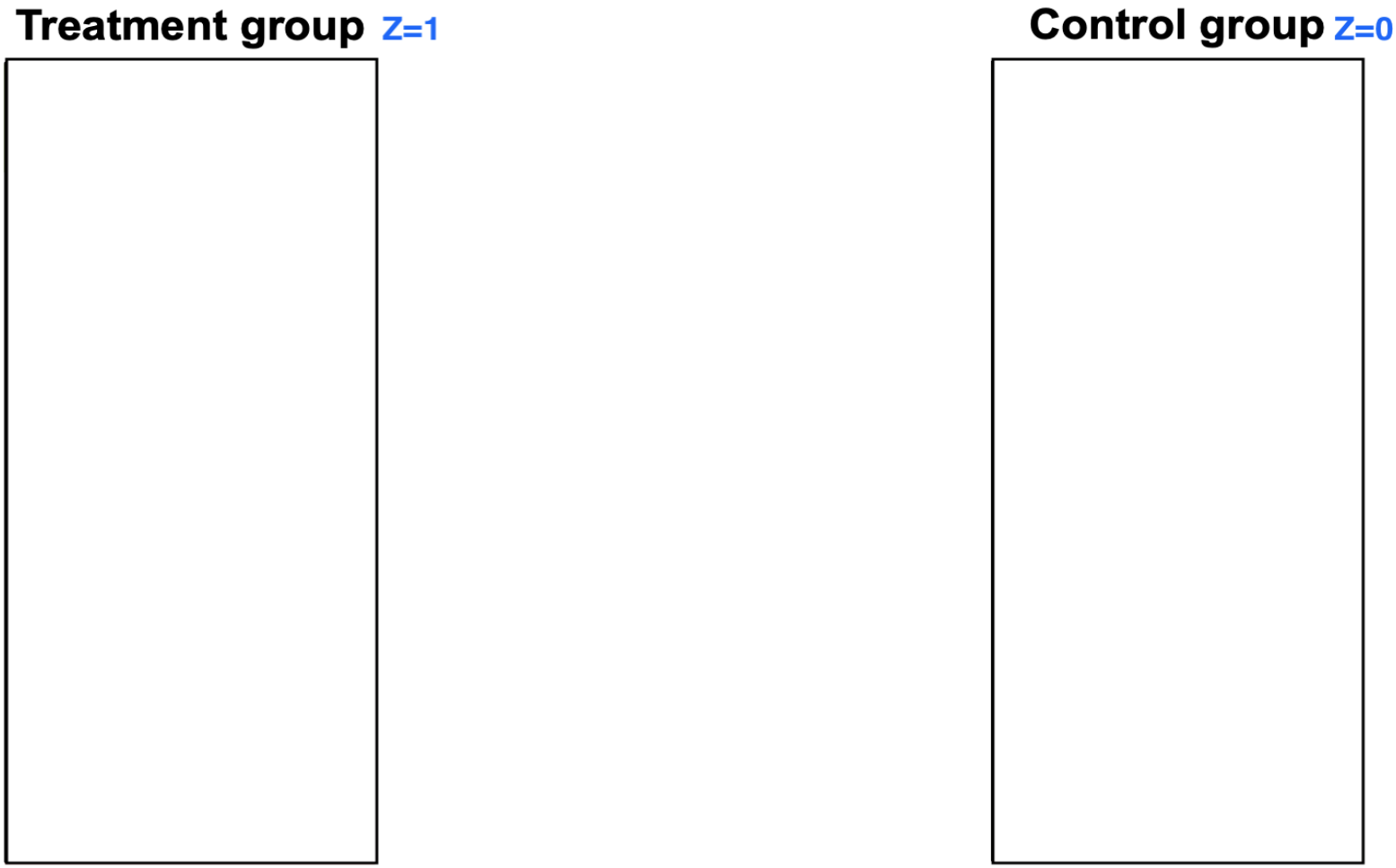
We use the same notation as in Lecture 1-2 and denote treatment *status* of individual i as

$$D_i = \begin{cases} 1 & \text{if she } \textit{receives} \text{ the treatment} \\ 0 & \text{otherwise} \end{cases}$$

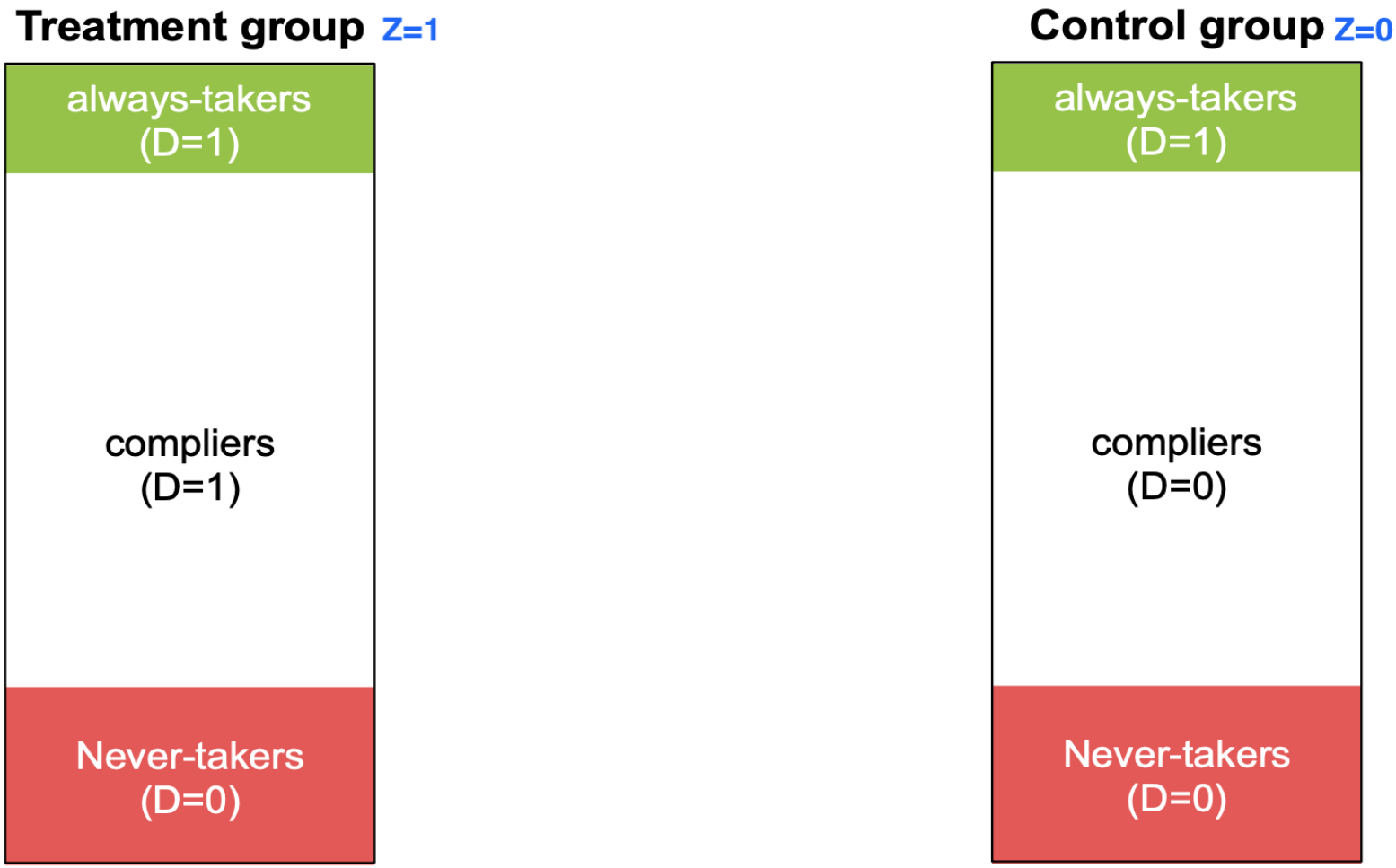
And the treatment *assignment* (randomization status) of i is:

$$Z_i = \begin{cases} 1 & \text{if she is } \textit{assigned} \text{ to treatment} \\ 0 & \text{if she is assigned to control} \end{cases}$$

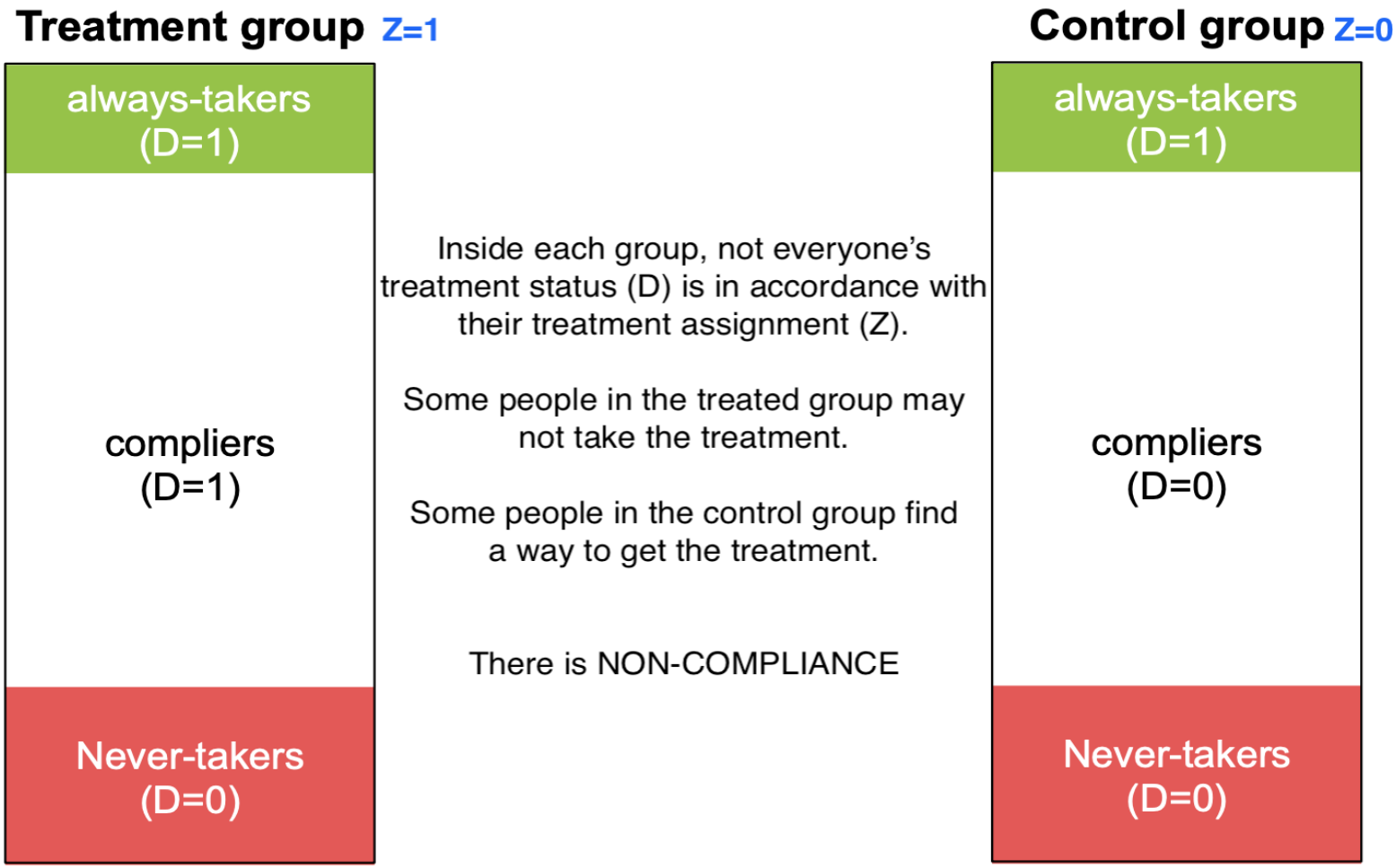
All units (e.g. individuals) in the treatment group have Z=1, i.e. they are assigned to treatment. All units in control have Z=0: they are not assigned to treatment.



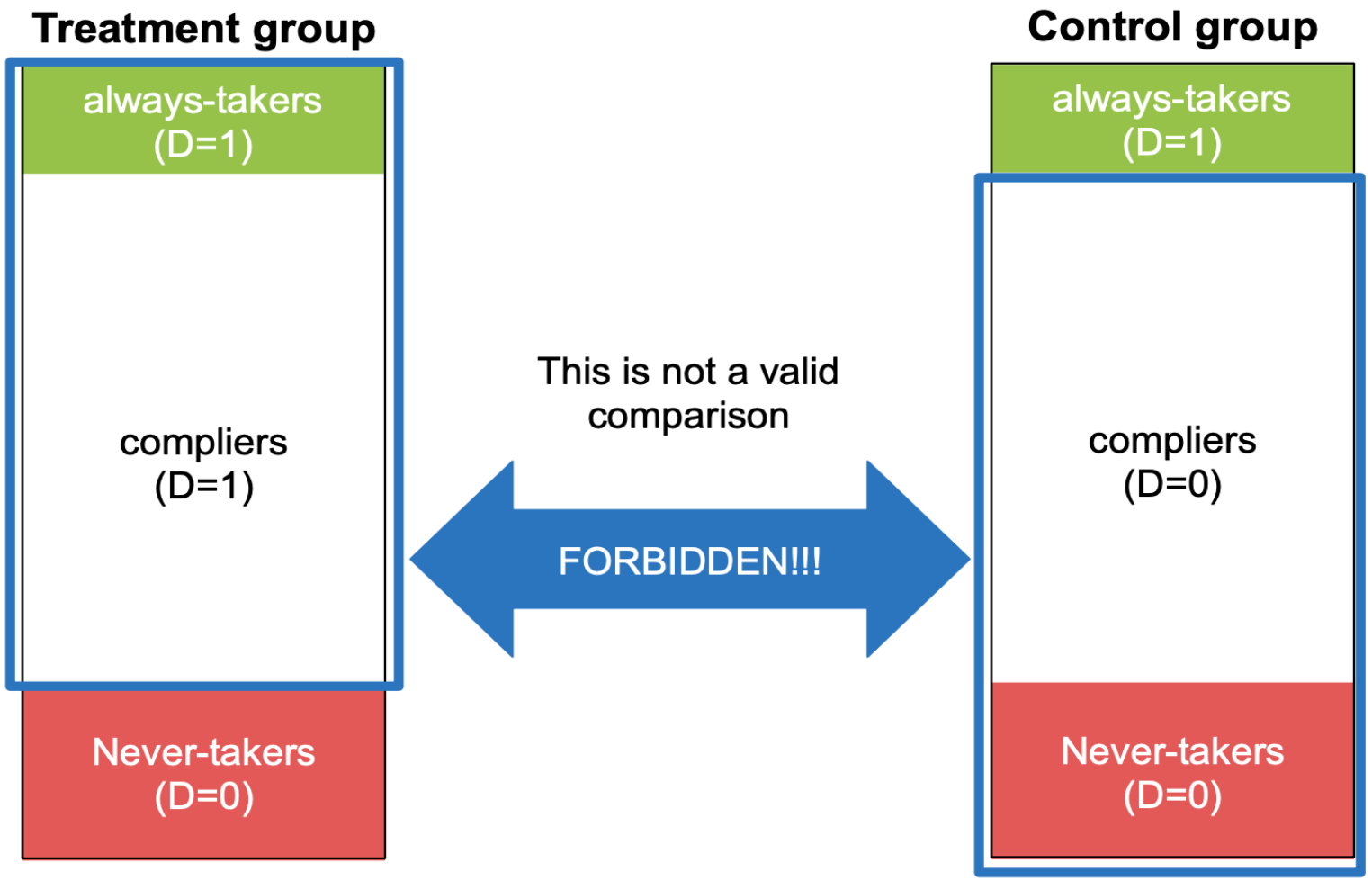
Inside the groups



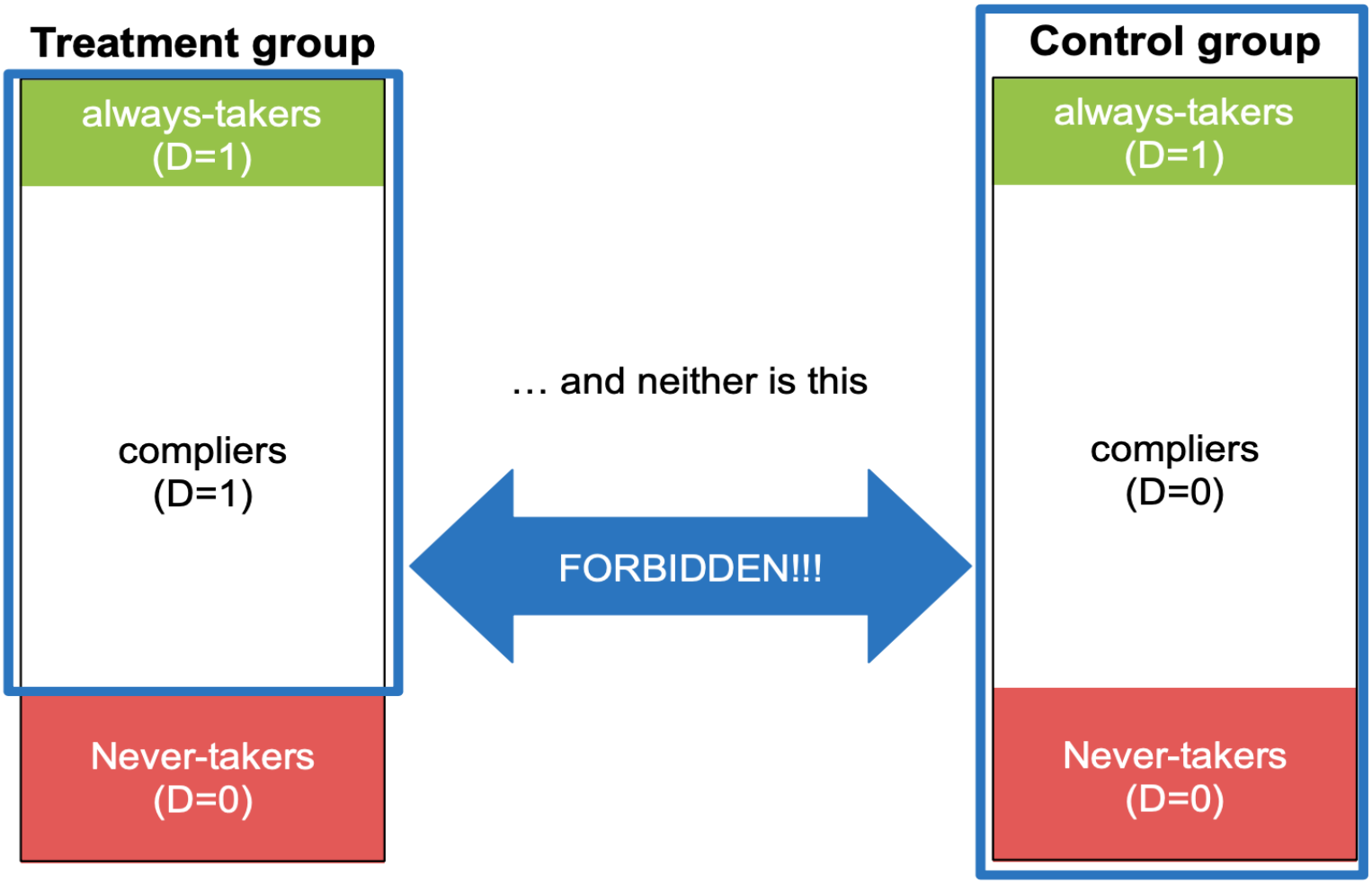
Inside the groups



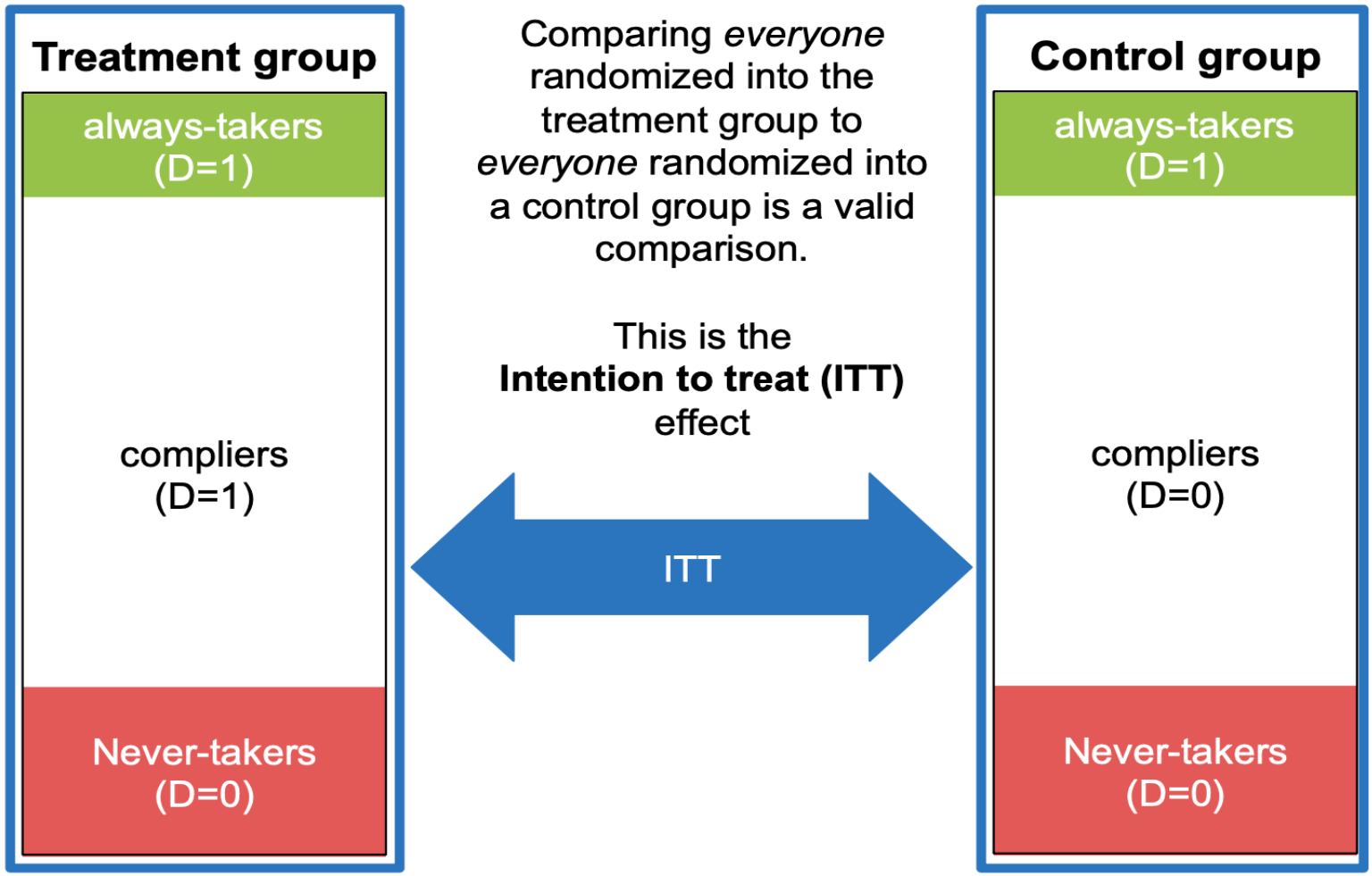
Invalid comparisons



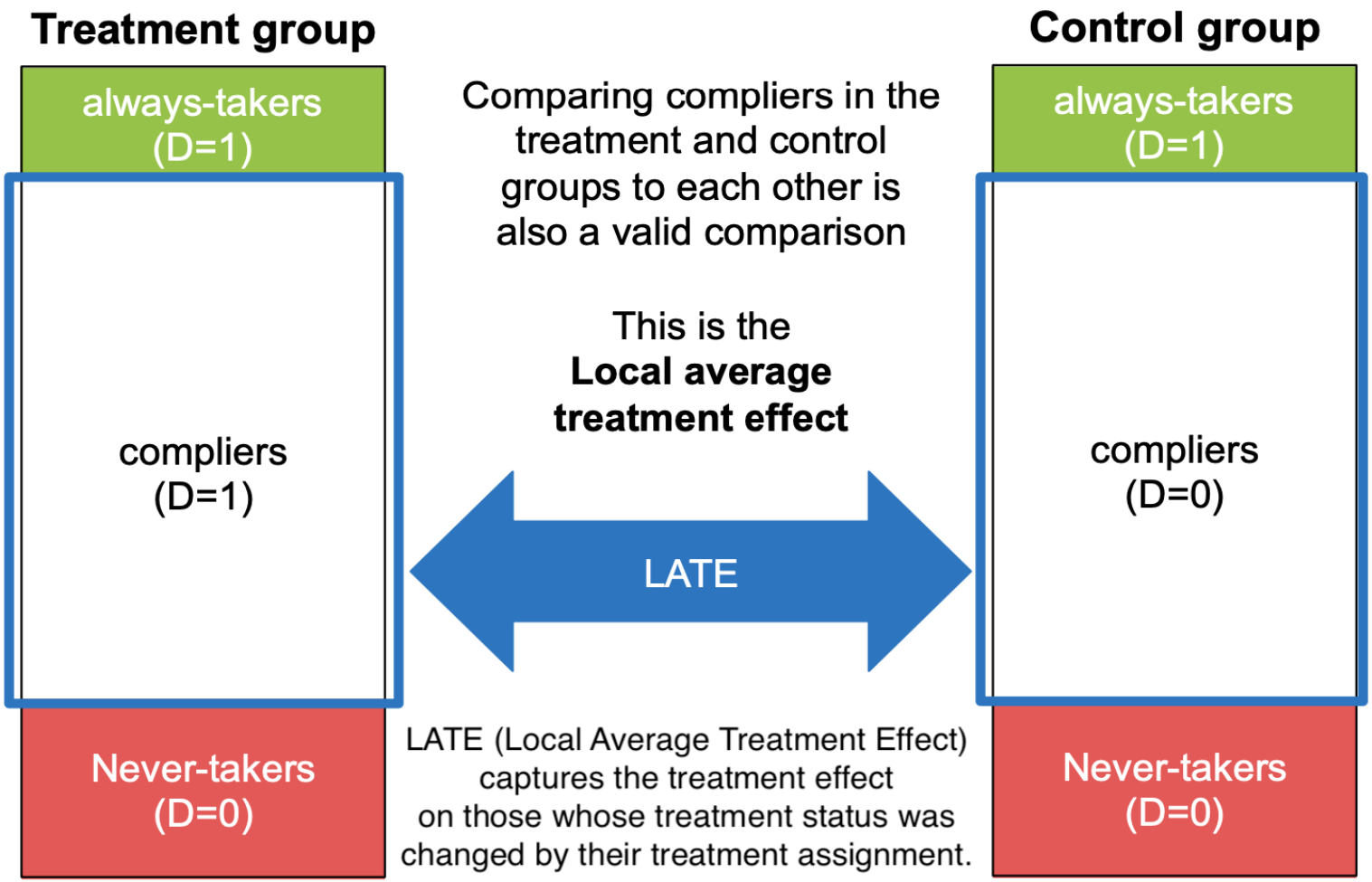
Invalid comparisons

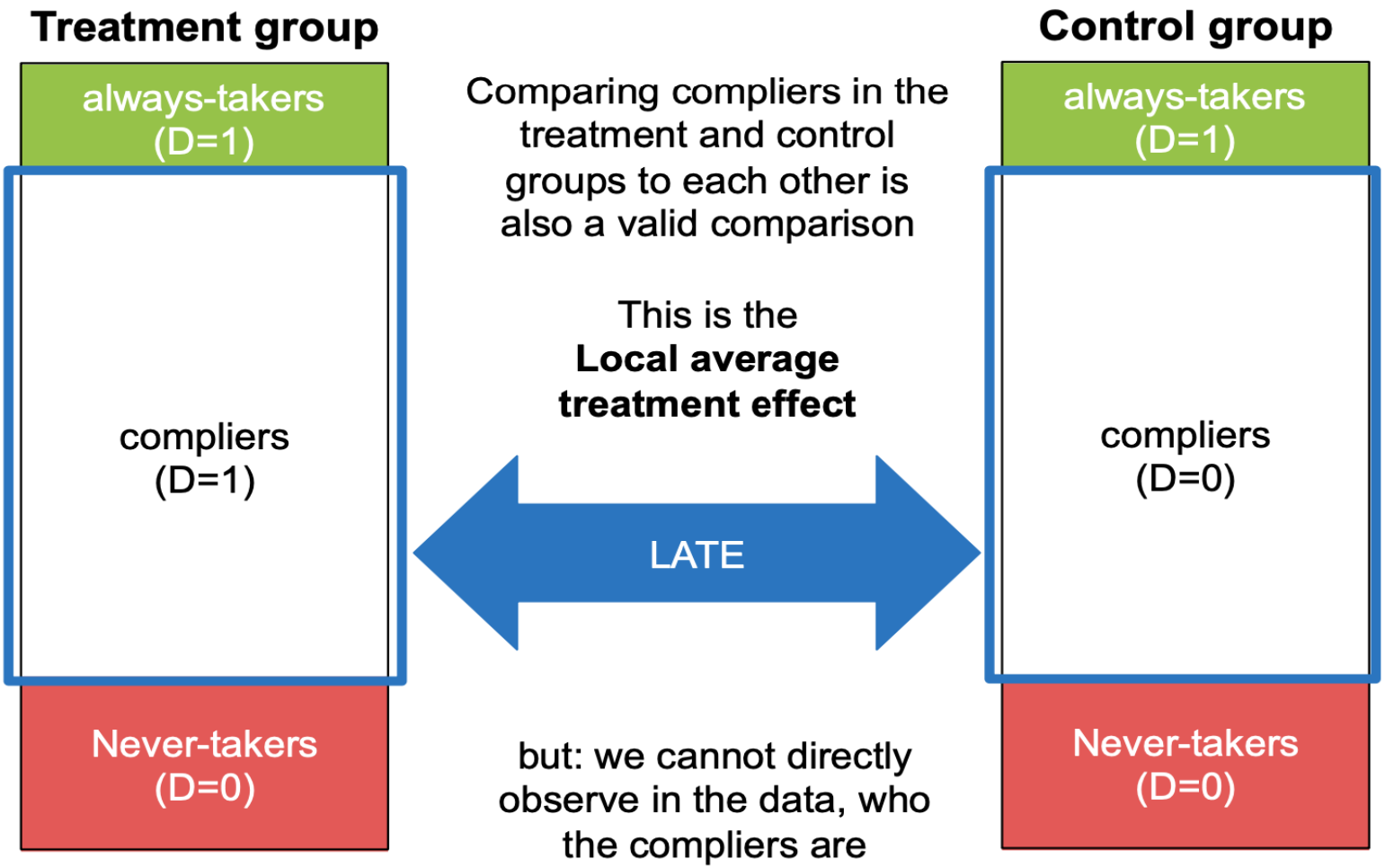


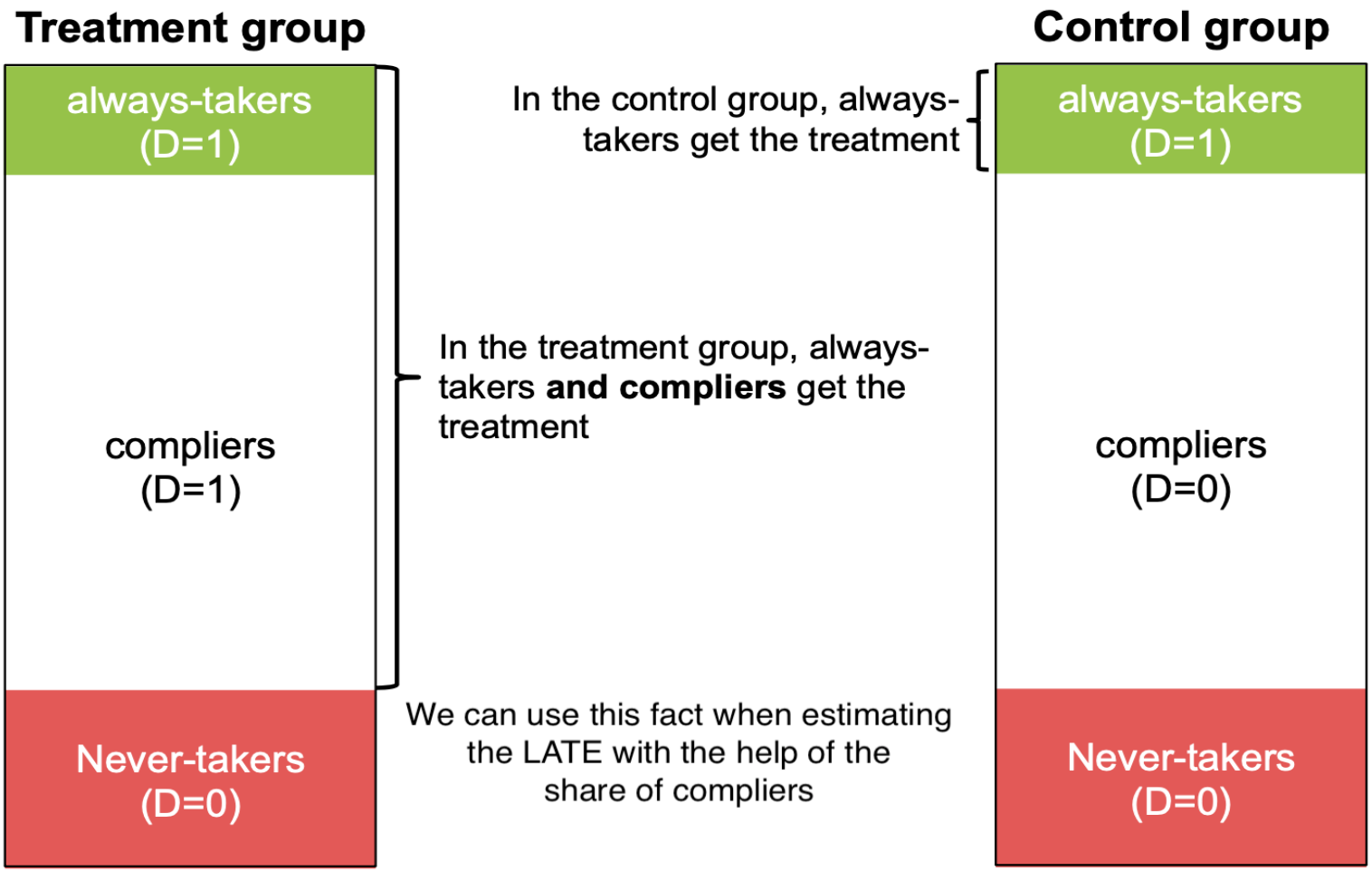
Valid comparison 1: ITT



Valid comparison 2: LATE




Treatment group
always-takers (D=1)
compliers (D=1)
Never-takers (D=0)
Comparing compliers in the treatment and control groups to each other is also a valid comparison
This is the Local average treatment effect
LATE
but: we cannot directly observe in the data, who the compliers are
Control group
always-takers (D=1)
compliers (D=0)
Never-takers (D=0)


Treatment group
Control group
always-takers (D=1)
In the control group, always-takers get the treatment
always-takers (D=1)
In the treatment group, always-takers **and compliers** get the treatment
compliers (D=1)
compliers (D=0)
We can use this fact when estimating the LATE with the help of the share of compliers
Never-takers (D=0)
Never-takers (D=0)

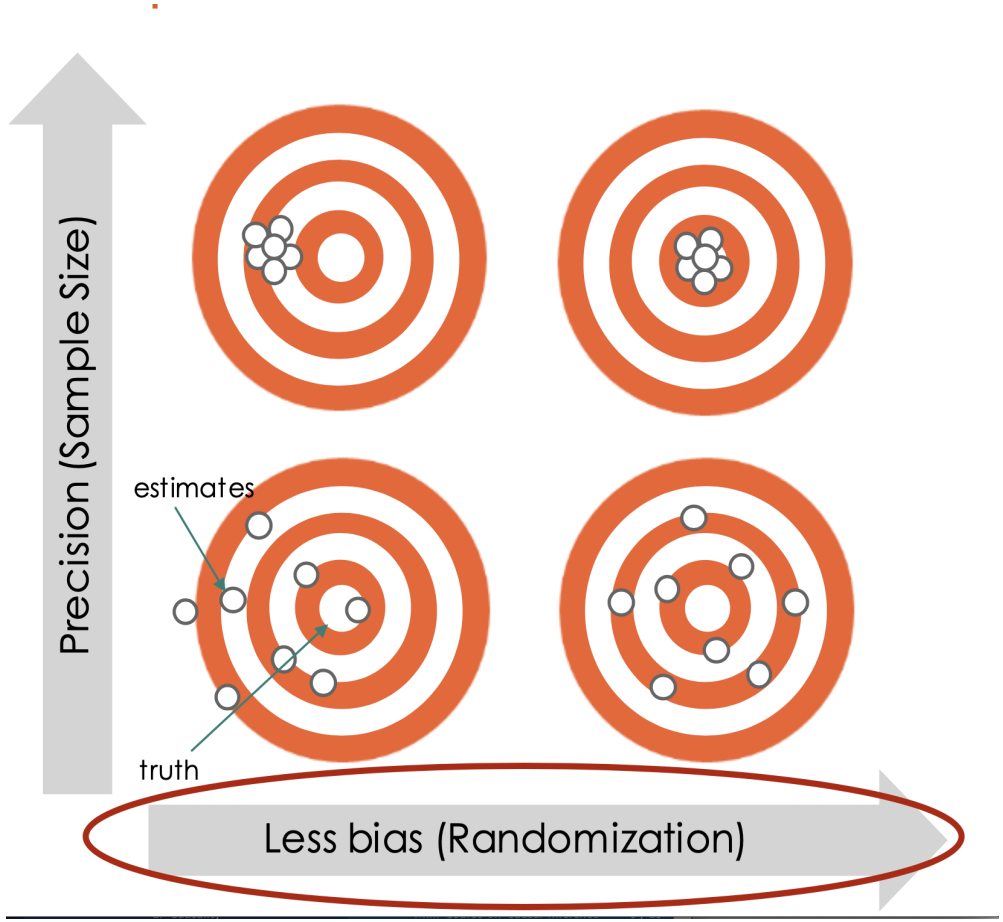
Under the assumption that the treatment and the control group are indeed comparable, and there is no differential selection into the groups we would expect

(i) same shares of always-takers across T and C groups

(ii) same shares of never-takers across T and C groups

We will now look more into comparability of T and C groups.

The balance check is a way to assess the risk of bias remaining despite the randomization.



Baseline balance

The first table in a paper that present the results from and RCT is usually a Balance table, where the researchers

- present an overview of the variables in the data, and
- check if there is **balance** on important variables at "baseline" = before the intervention began.

  $\Rightarrow$ In other words: are the treatment and control groups comparable and similar on key characteristics?

Baseline balance table

TABLE 1A—BASELINE SUMMARY STATISTICS

	Control group			Treatment − control	
	Obs. (1)	Mean (2)	SD (3)	Coeff. (4)	*p*-value (5)
Household composition					
Number members	1,220	5.038	(1.666)	0.095	0.303
Number adults (>=16 years old)	1,220	3.439	(1.466)	−0.011	0.873
Number children (<16 years old)	1,220	1.599	(1.228)	0.104	0.098
Male head	1,216	0.907	(0.290)	−0.012	0.381
Head's age	1,216	41.150	(10.839)	−0.243	0.676
Head with no education	1,216	0.370	(0.483)	−0.008	0.787
Access to credit					
Loan from Spandana	1,213	0.000	(0.000)	0.007	0.195
Loan from other MFI	1,213	0.011	(0.103)	0.007	0.453
Loan from a bank	1,213	0.036	(0.187)	0.001	0.859
Informal loan	1,213	0.632	(0.482)	0.002	0.958
Any type of loan	1,213	0.680	(0.467)	0.002	0.942
Amount borrowed from (in Rs)					
Spandana	1,213	0	(0.000)	69	0.192
Other MFI	1,213	201	(2,742)	170	0.568
Bank	1,213	7,438	(173,268)	−5,420	0.279
Informal loan	1,213	28,460	(65,312)	−570	0.856
Total	1,213	37,892	(191,292)	−5,879	0.343

Mean household size in control group is 5.038 people. The mean in the treatment group is 5.038+0.095= 5.133

P-values of difference: we start worrying that groups are not comparable if p goes below < 0.1

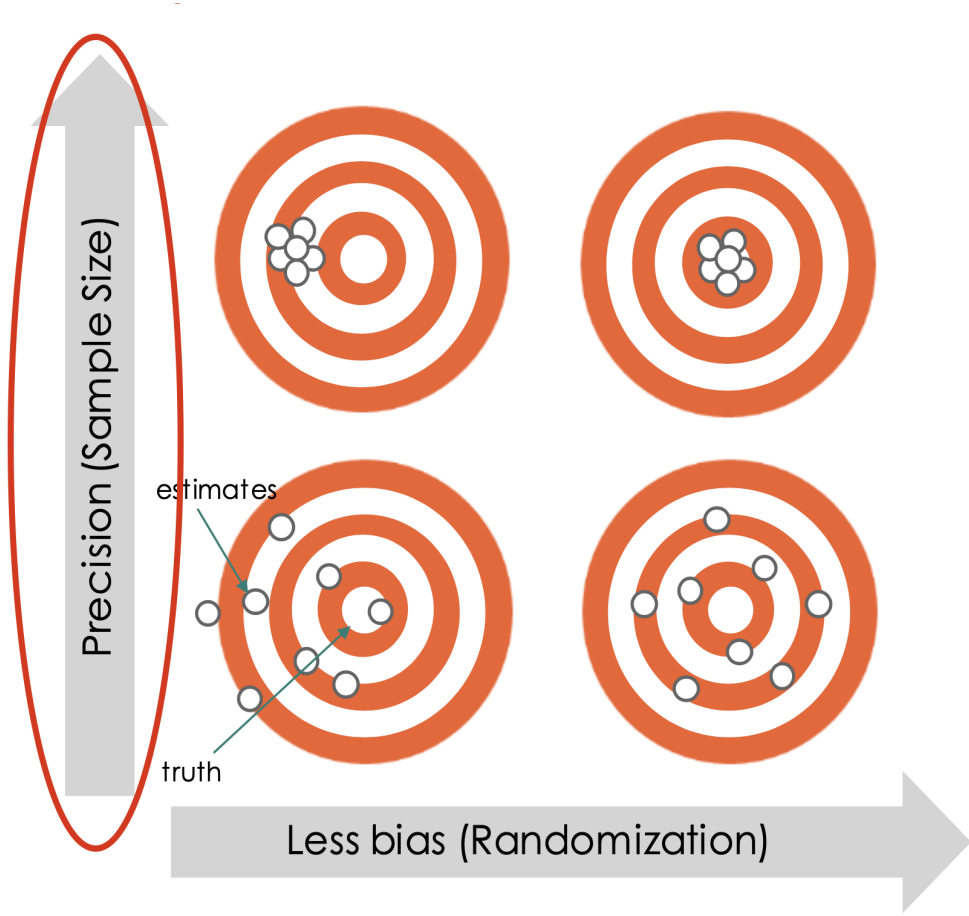
Randomized experiments - risks and limitations

- Most important: show that the *outcome variable* of interest is balanced at baseline, (if it can be measured already at baseline).

Baseline balance table

				difference Treat-Control	p-value of difference
Self-employment activities					
Number of activities	1,220	0.320	(0.682)	−0.019	0.579
Number of activities managed by women	1,220	0.145	(0.400)	−0.007	0.750
Share of HH activities managed by women	295	0.488	(0.482)	−0.006	0.904
Businesses					
Revenue/month (Rs)	295	15,991	(53,489)	4,501	0.539
Expenses/month (Rs)	295	3,617	(26,144)	641	0.751
Investment/month (Rs)	295	385	(3,157)	14	0.959
Employment (employees)	295	0.169	(0.828)	0.255	0.148
Self-employment (hours per week)	295	76.315	(66.054)	−4.587	0.414
Businesses (all households)					
Revenue/month (Rs)	1,220	3,867	(27,147)	904	0.626
Expenses/month (Rs)	1,220	875	(12,933)	116	0.812
Investment/month (Rs)	1,220	93	(1,559)	−0.098	0.999
Employment (employees)	1,220	0.041	(0.413)	0.057	0.166
Self-employment (hours per week)	1,220	18.453	(46.054)	−1.801	0.400
Consumption (per household per month)					
Total consumption (Rs)	1,220	4,888	(4,074)	270	0.232
Nondurables consumption (Rs)	1,220	4,735	(3,840)	252	0.235
Durables consumption (Rs)	1,220	154	(585)	18	0.531
Asset index	1,220	1.941	(0.829)	0.027	0.669

We will now talk about considerations related to precision of our estimate



Statistical power

- Statistical power: how likely are we to conclude that a treatment has an impact, when it truly has an impact? Avoiding Type 2-error.
- Especially in randomized field experiments when the researcher is constrained in number of units that can be included, the resulting sample size is often too small. Constraints are caused by
  - budget - in some cases treatment can only be offered to a given number of people.
  - design/outcome: for some outcomes, randomizing at "cluster" level makes more sense than individual randomization.
- When the sample for some of the analysis depends on take-up, the risk of being underpowered is even higher.

Power: main ingredients

Power is affected by:

- Effect size (& take-up rate)
- Sample size (&number of clusters)
- Variance
- Proportion in sample in Treatment vs. Control
- Desired significance level (standard: 5%)

more on Power

Effect size and take up

The smaller the effect size that researchers want to be able to detect $\Rightarrow$ the larger the sample needed for a given level of significance.

- If the treatment is something where there is non-compliance and the take-up rate is low, a larger sample is needed than with full compliance.
- You can think of it as the average effect size among those assigned to treatment (ITT) being diluted.
- For more on this, see `https://blogs.worldbank.org/impactevaluations/power-calculations-101-dealing-with-incomplete-take-up`

When reading a paper where the take-up of treatment is low, check: did the authors account for this?

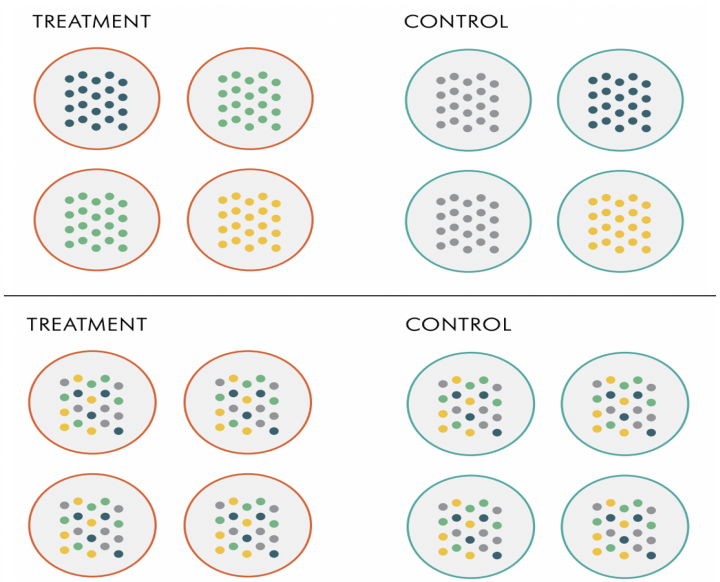
Sample size and clusters

A larger sample ⇒ higher power.

- If treatment is randomized at the individual level, including more individuals in the randomization ⇒ additional independent observations & More precision.
- However, often, treatment is randomized at the "cluster" level: e.g. schools, districts, neighborhoods, and the individuals within the cluster are all treated.
- If treatment is clustered but we are measuring individual responses, we need to take into account the correlation between individuals within same "cluster".
- Usually the number of clusters is the key determinant of power, not the number of people per cluster.

Clusters

- **Extreme case 1:** Here, all 20 individuals in each of the 4 clustes are identical: this sample gives us the same power as we would have in an RCT with only 8 individuals.
- **Extreme case 2**: Here, there is no correlation between individuals within a cluster, and we have same power as in individual randomization w. 80 individuals each in T and C.



Case study; Miracle of Microfinance

In the aforementioned study by Banerjee &Duflo, the researchers wanted to estimate the effect of microfinance services on various firm and household outcomes.

- the initial power calculations were performed when researchers thought 80% of eligible households would become clients.
- In fact, the proportion reached only 18 percent in 18 months.
- $\Rightarrow$ in hindsight, many more neighborhoods would have been needed. This is not something that could be addressed ex post.

Case study; Miracle of Microfinance

- Results show weak effects of Microfinance on various welfare outcomes.
  - small point estimates (suggesting smaller "effect" than expected)
  - Statistically insignificant estimates
- Why are the estimates so small?
- This could be either because the true effect is small, or because the sample is somehow not representative, and by chance the effect in *this* sample is small. Recall that the smaller the coefficient, the larger a sample is required to obtained statistical power.
- The authors' "solution": *"Fortunately, subsequent evaluations of microfinance programs [with larger samples] find a very similar set of results (and non-results), suggesting that these outcomes are not the artifact of samples that are too small or of a very non-representative set of clients."*

Design considerations

Suppose you are a researcher who wants to estimate the effect of microfinance loans on business profits of small businesses, by comparing (a) small business owners who have/use microfinance to those (b) who do not. But several ways to do this:

- Naive approach: Comparing current borrowers to non-borrowers? $\Rightarrow$ Not good: likely to be affected by selection

Design considerations: Level of randomization

- Design an experiment with random assignment that solves the selection problem, But several ways to do this too!
  - Village level: Randomly assign microfinance to some villages and not to others, and comparing the population of the villages? Now the selection problem is solved by random assignment. But what are we picking up?
  - Individual level 1: Randomly assigning some *individuals* to take a loan and others not to take a loan? But we cannot risk force people to take a loan, so risk of low take-up, and selection.
  - Individual level 2: Focusing on applicants for a loan who were marginally rejected, and assigning some of them randomly to a loan, while control group are not offered loans? This was done in some papers. Ensures that entire sample is interested in a loan, but limit external validity of the results.

Design considerations and power*

The design of the experiment can also affect compliance and thereby statistical power.

- Example 2: we want to evaluate a business training program for small business owners.
  - Approach 1: an encouragement design, where randomly selected clients are asked whether they want to participate in the program, and they could choose whether or not to do it. The evaluation would then compare those invited to those who were not invited.
  - Approach 2: an oversubscription design, where clients are asked to apply, and the program is then randomized among applicants. The take-up of the program in the second design would presumably be much larger than that in first design.

Summary

We have discussed

1. Causal comparisons and non-compliance
2. Issues related to **bias**:
   - Balance checks
3. Issues related to **precision**:
   - Statistical Power, sample size and take-up
4. Design considerations

Useful links

For more on reading Baseline tables and other tables in RCT papers, we highly recommend to watch the following video with Josh Angrist:

- `https://youtu.be/s-_3s30Meqs`

For more information and tools to calculate power, see

- Optimal design free software for PC `http://hlmsoft.net/od/`
- `https://www.povertyactionlab.org/resource/quick-guide-power-calculations`

Power: equation

$$MDE = (t_{(1-\kappa)} + t_{\alpha}) \times \sqrt{\frac{1}{P(1-P)}} \times \sqrt{\frac{\sigma^2}{n}}$$

- MDE= Effect size (Minimum detectable)
- $t_{(1-\kappa)}$=power; t_{α}=significance level
- P=share of sample in the treatment group
- σ^2= variance
- n= sample size

Back

Power: equation with clusters

$$\frac{MDE}{\sqrt{1+\rho(m-1)}} = (t_{(1-\kappa)} + t_{\alpha}) \times \sqrt{\frac{1}{P(1-P)}} \times \sqrt{\frac{\sigma^2}{n}}$$

- MDE= Effect size (Minimum detectable)
- ρ = Intra cluster correlation (picking up how similar the units within each cluster are to each other)
- m=average cluster size, e.g. if a cluster is a household, and average hh size in our sample is 5, m=5.

Back

Power - idea (with H_0=No effect)*

		The Truth (Based on Entire Population)	
		Nothing Is There (H_0 Is True)	Something Is There (H_0 Is False)
Your Conclusion (Based on Your Sample)	I Don't See Anything (Nonsignificant)	Right!	Wrong (Type II Error)
	I See Something (Significant)	Wrong (Type I Error)	Right!

A risk in randomized experiment: too few observations (units) leads to Type 2 error: study is underpowered. Back