Counting Defiers: A Design-Based Model of an Experiment Can Reveal Evidence Beyond the Average Effect

Neil Christy and Amanda Kowalski
University of Michigan

Thanks to Jann Spiess for extensive regular feedback and to Charles Manski, Toru Kitagawa, Aleksey Tetenov, and Donald Rubin for encouraging us to use statistical decision theory and teaching us about it. Thanks also to Guido Imbens.

Paper on arXiv: https://arxiv.org/abs/2412.16352

You want pregnant women to take up a desired health behavior.

You consider a financial intervention.

But you worry it could backfire for some people.

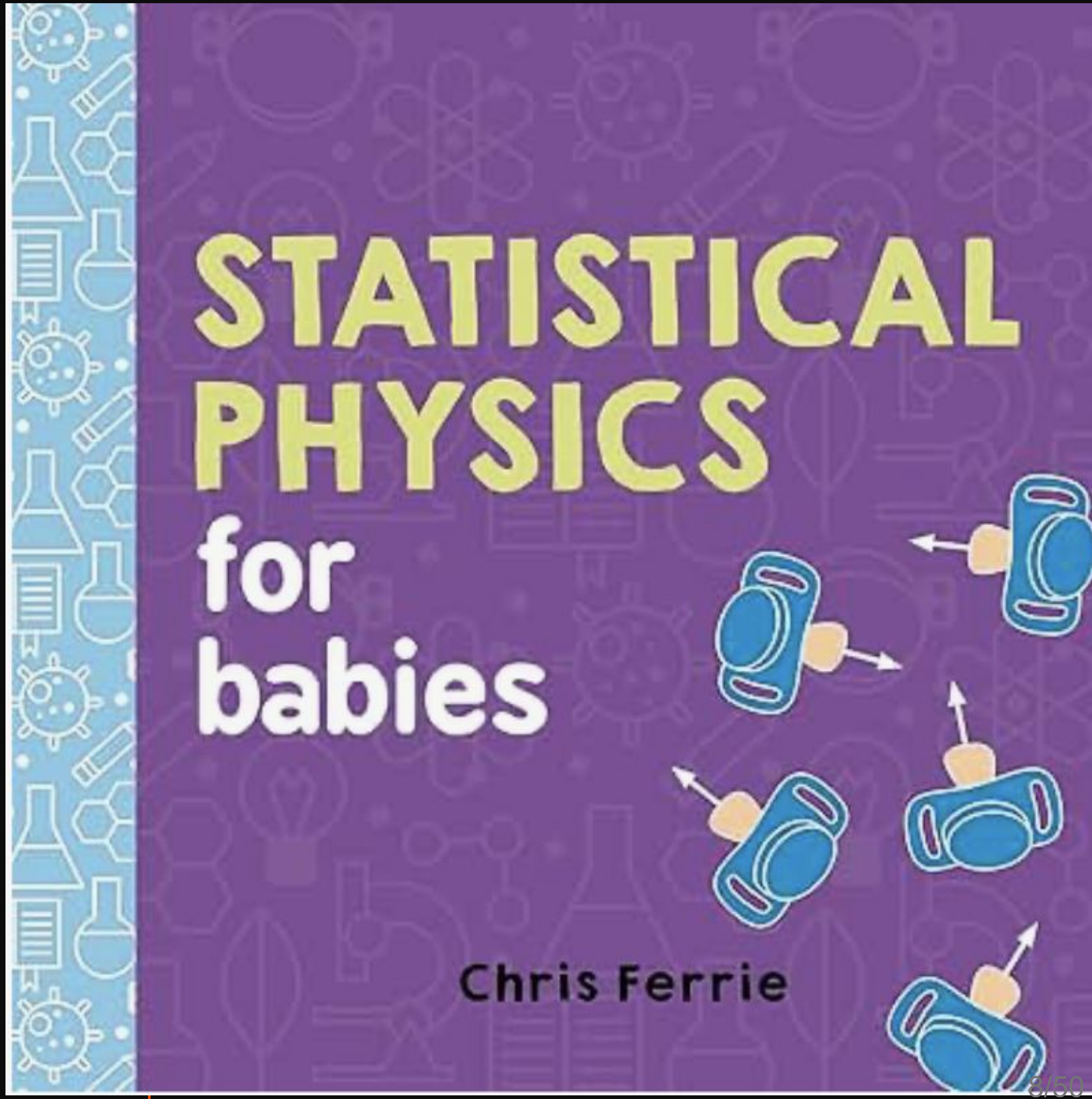
You run a randomized experiment with six people, three in intervention.

Two take up in intervention, and one takes up in control.

Does the intervention backfire for any of them?

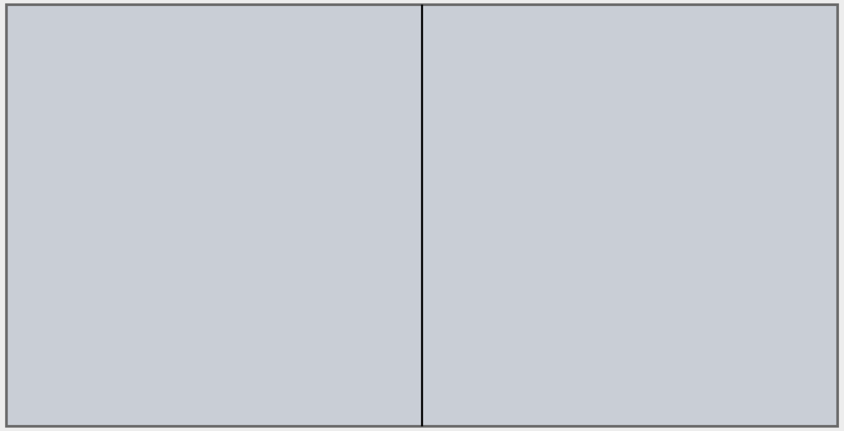
I'll help you decide, demonstrating that

A design-based model of an experiment can reveal evidence beyond the average effect.

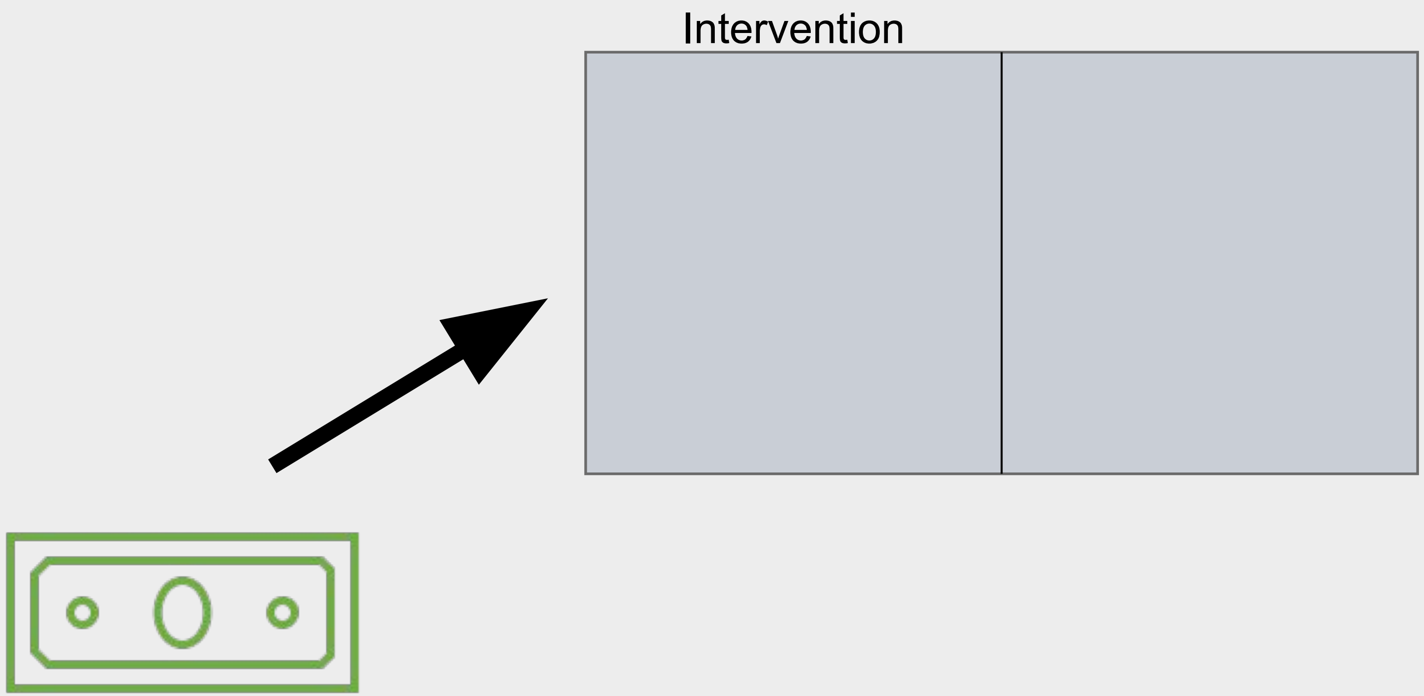


Pictured book and style inspiration for visualizations: Ferrie 2016

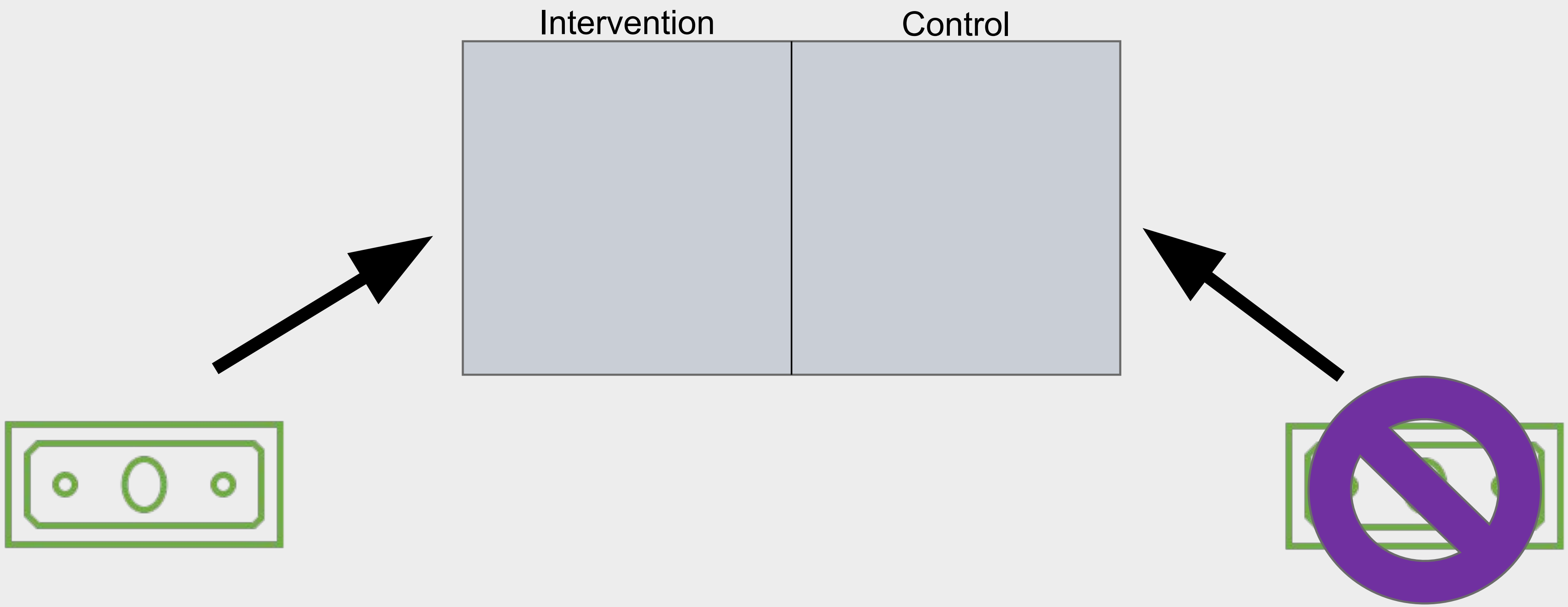
This is your randomized experiment.



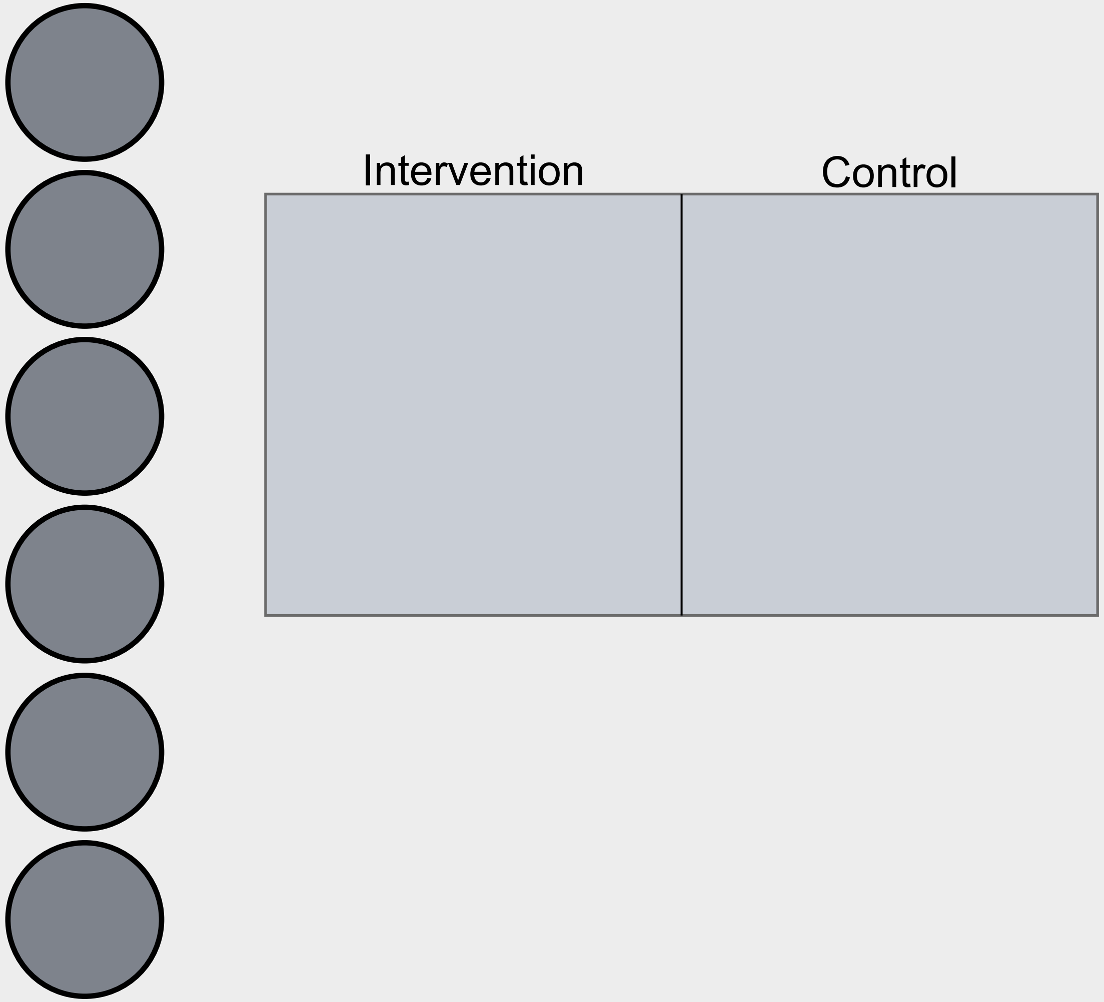
It has an **intervention** on the left,



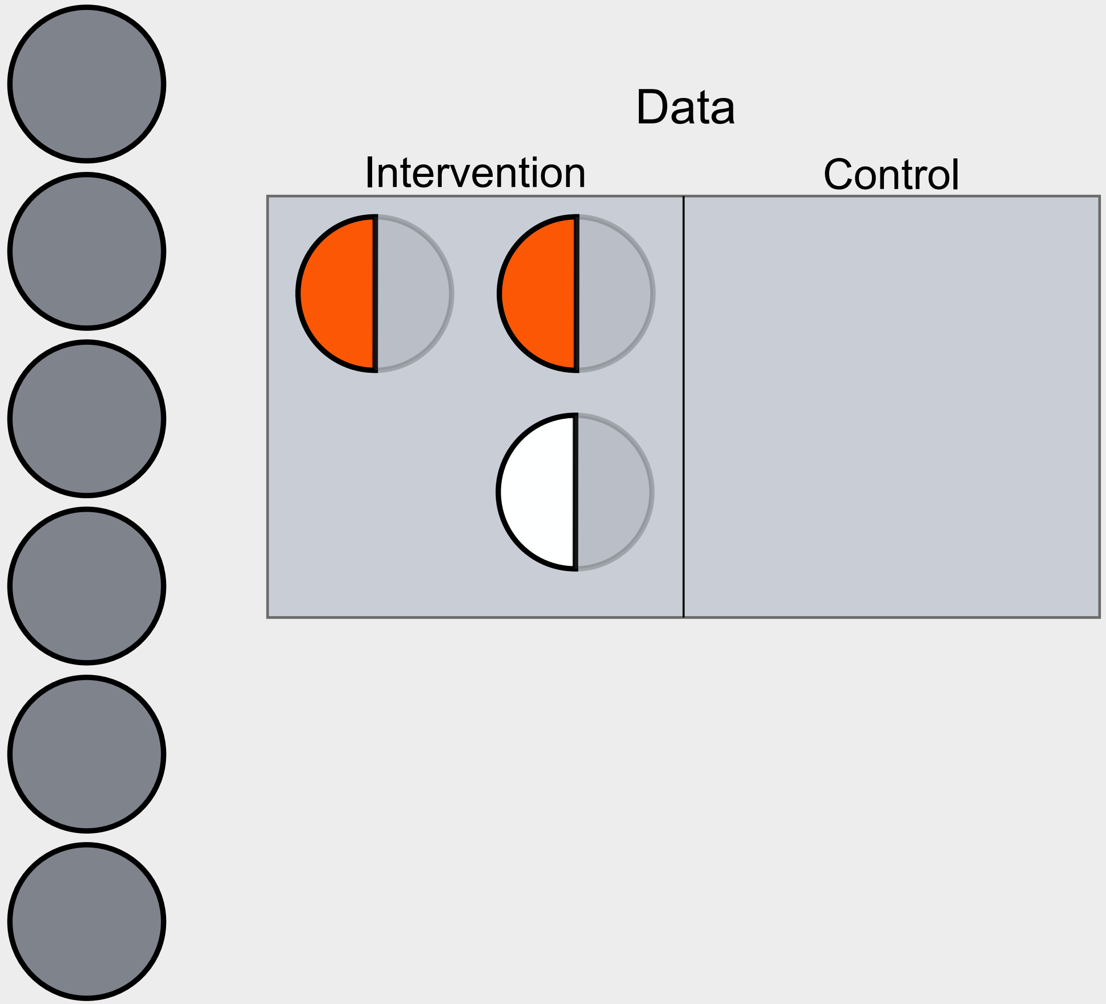
It has an **intervention** on the left, and a **control** on the right.



There are six people,

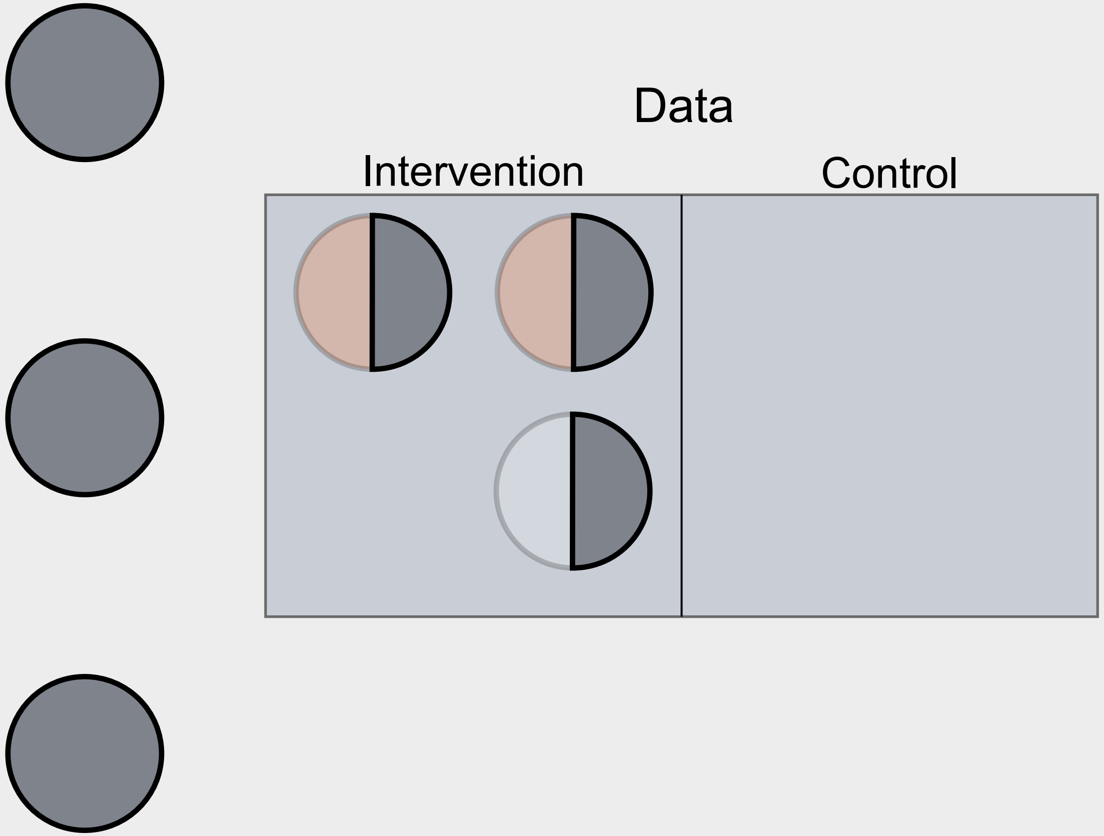


There are six people, and three are randomly assigned intervention.
You observe takeup and no takeup in intervention.



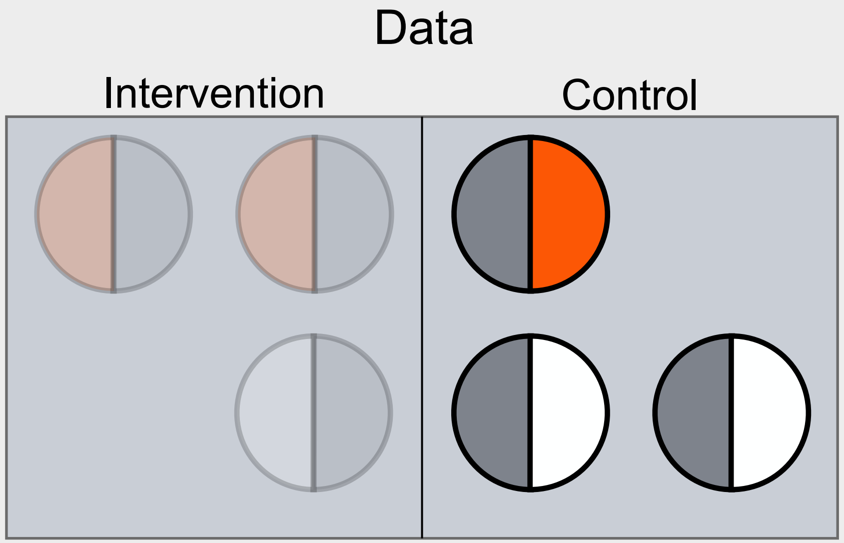
Intervention: Takeup No Takeup Unobserved

But the **potential outcome** in control remains unobserved (Rubin 1974).



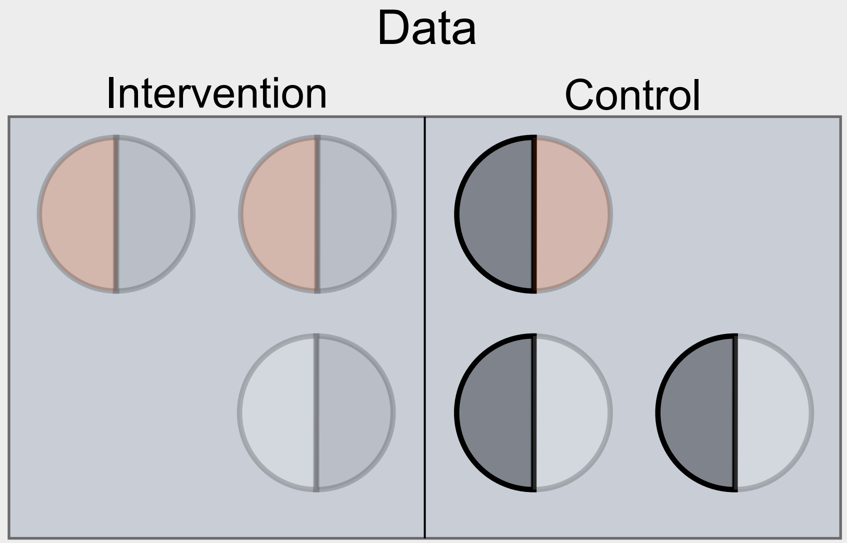
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

For the people in control, you observe takeup and no takeup in control.



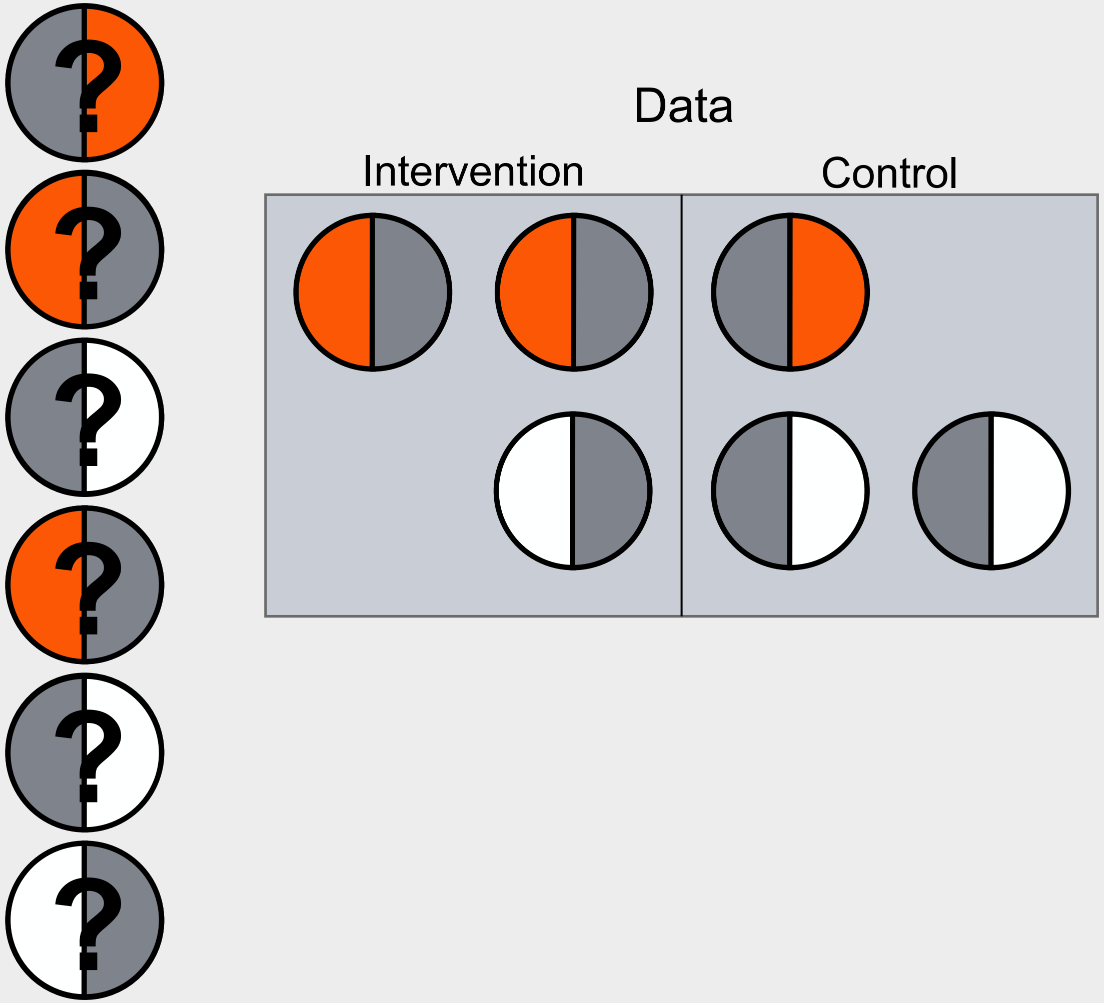
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

But the potential outcome in intervention remains unobserved.



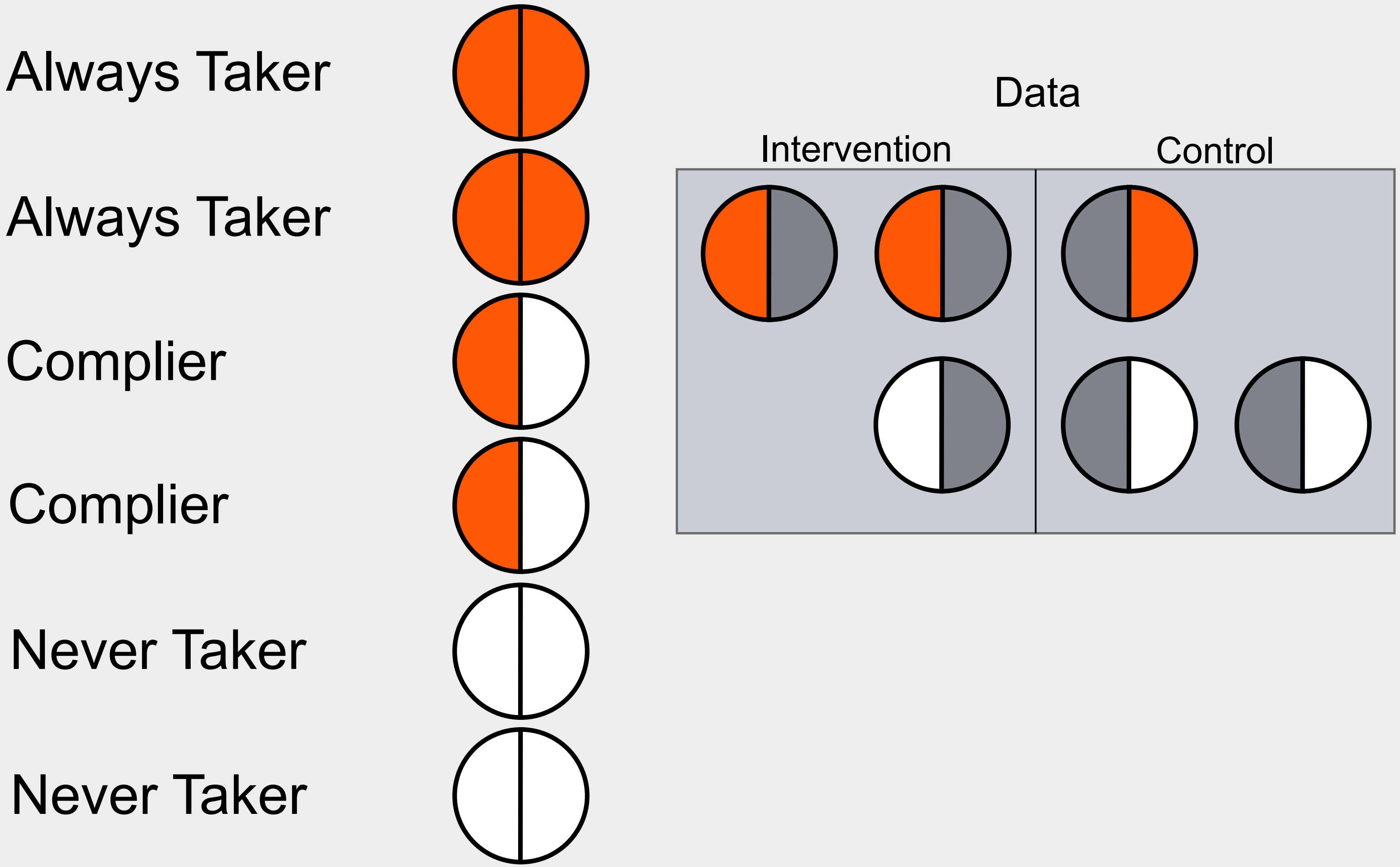
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

What is your best guess of the unobserved potential outcomes that generated the data?
What is the joint distribution of potential outcomes?



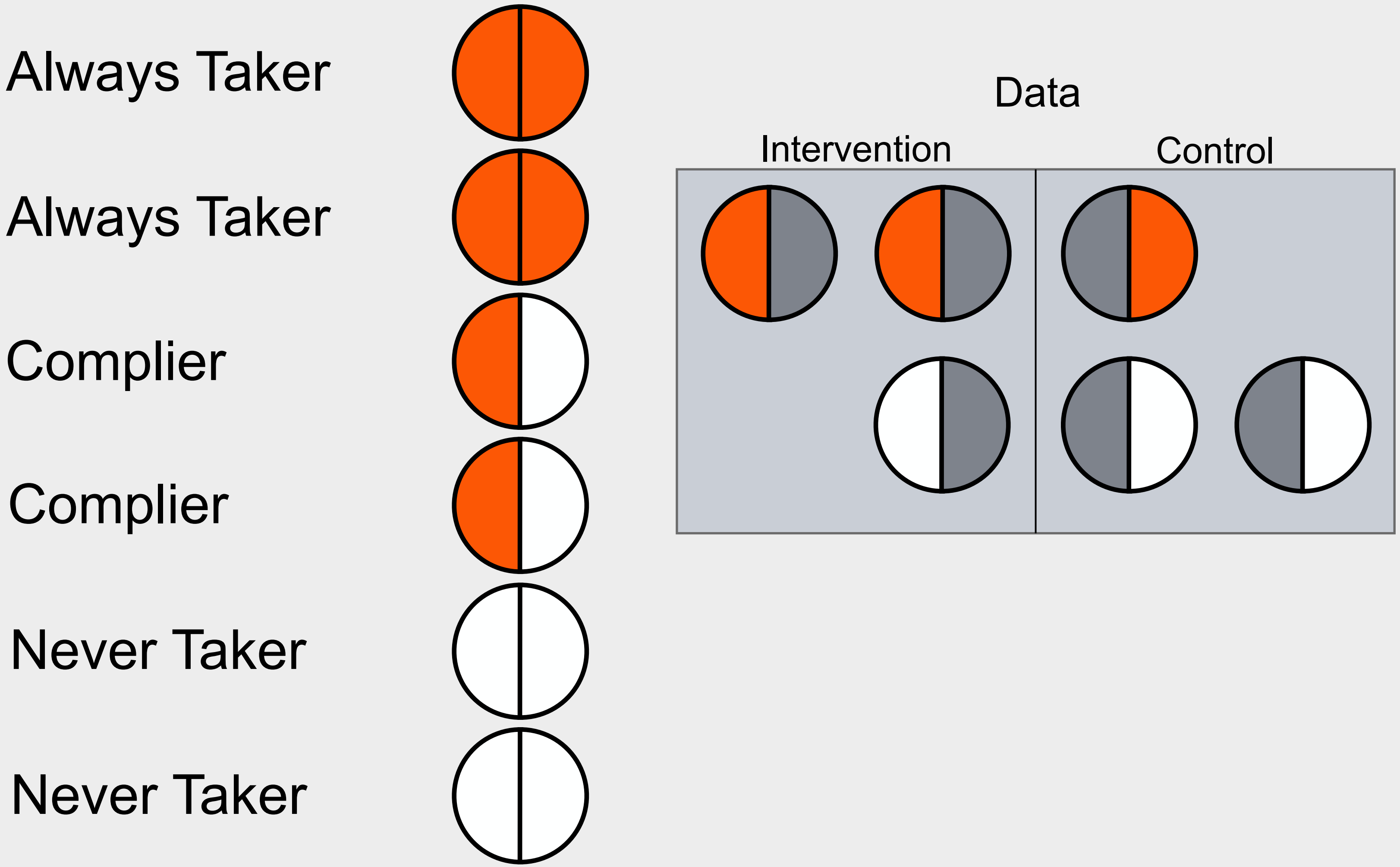
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

Here is one possible answer.



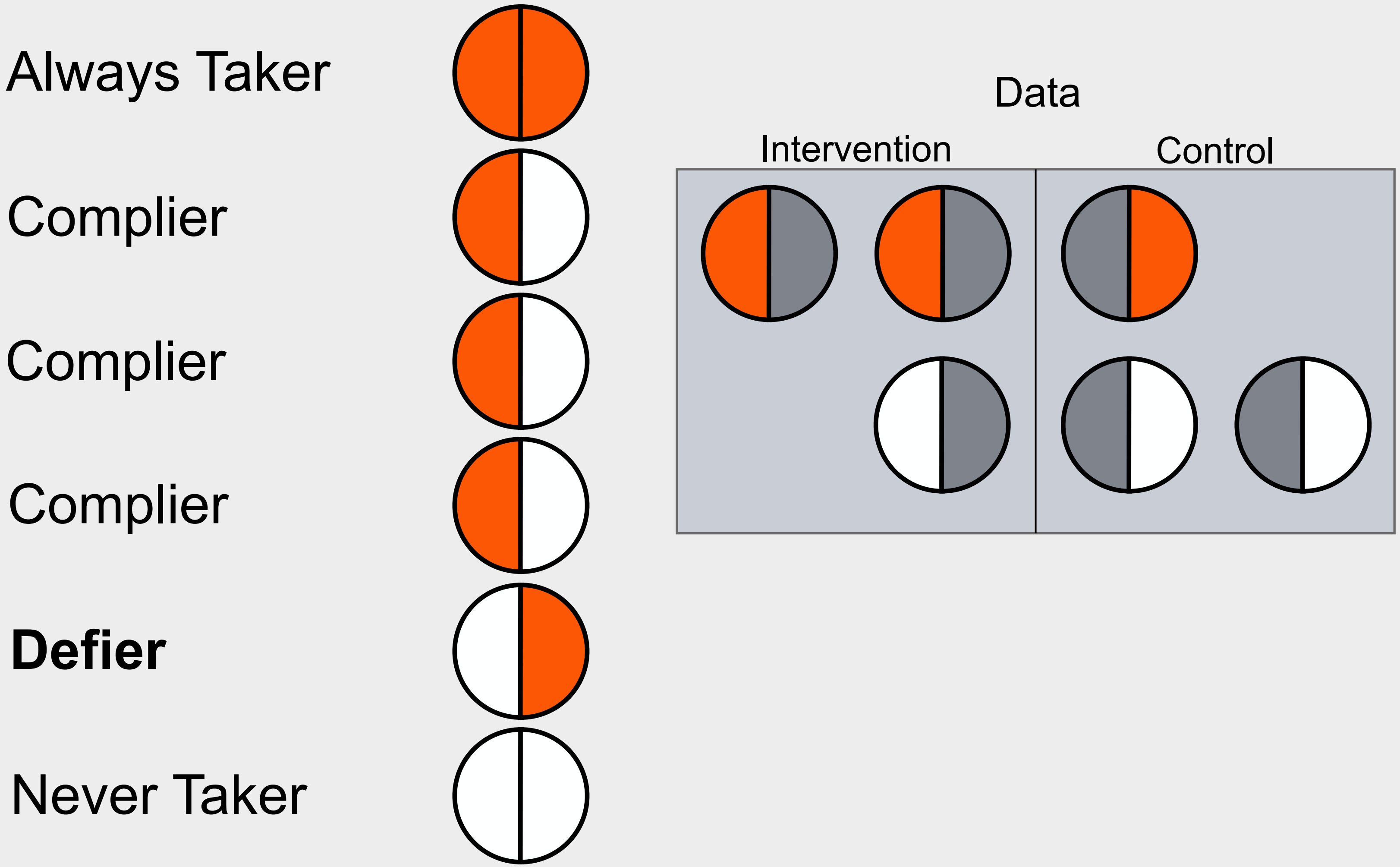
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

This answer satisfies **monotonicity** (Imbens and Angrist 1994).



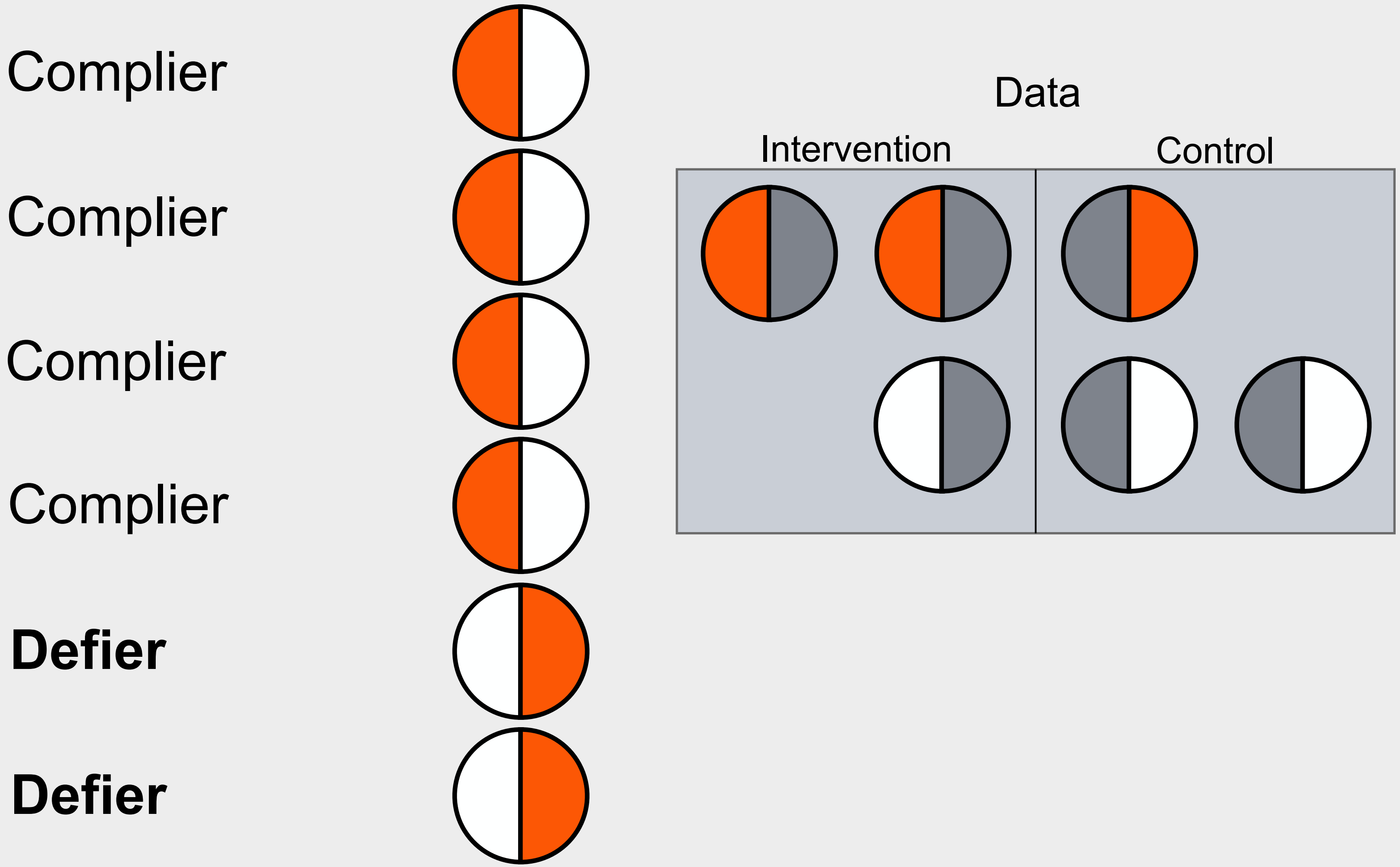
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

Here is another answer with a **defier** (Angrist, Imbens, and Rubin 1996).



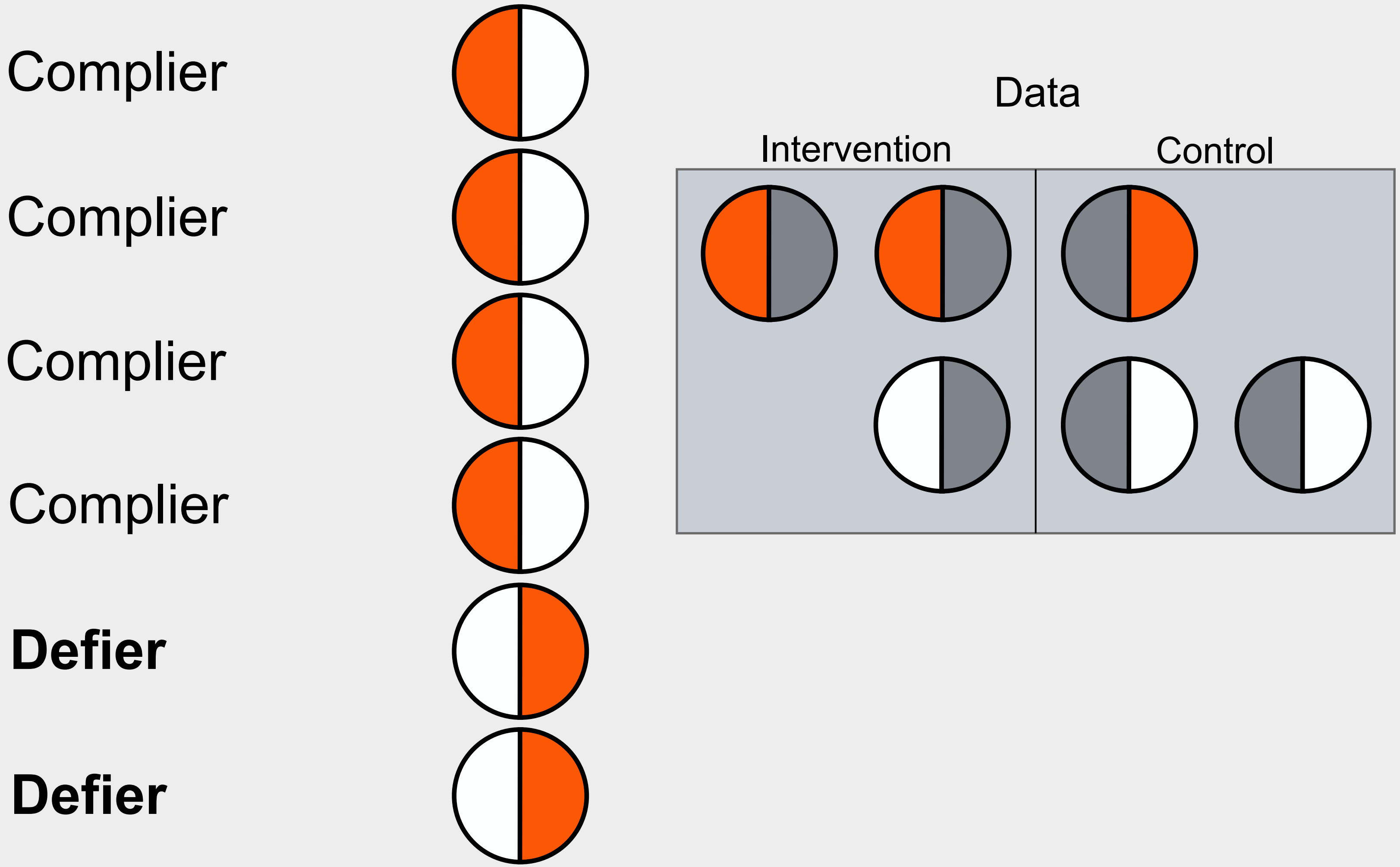
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

Here is another answer with **two defiers.**



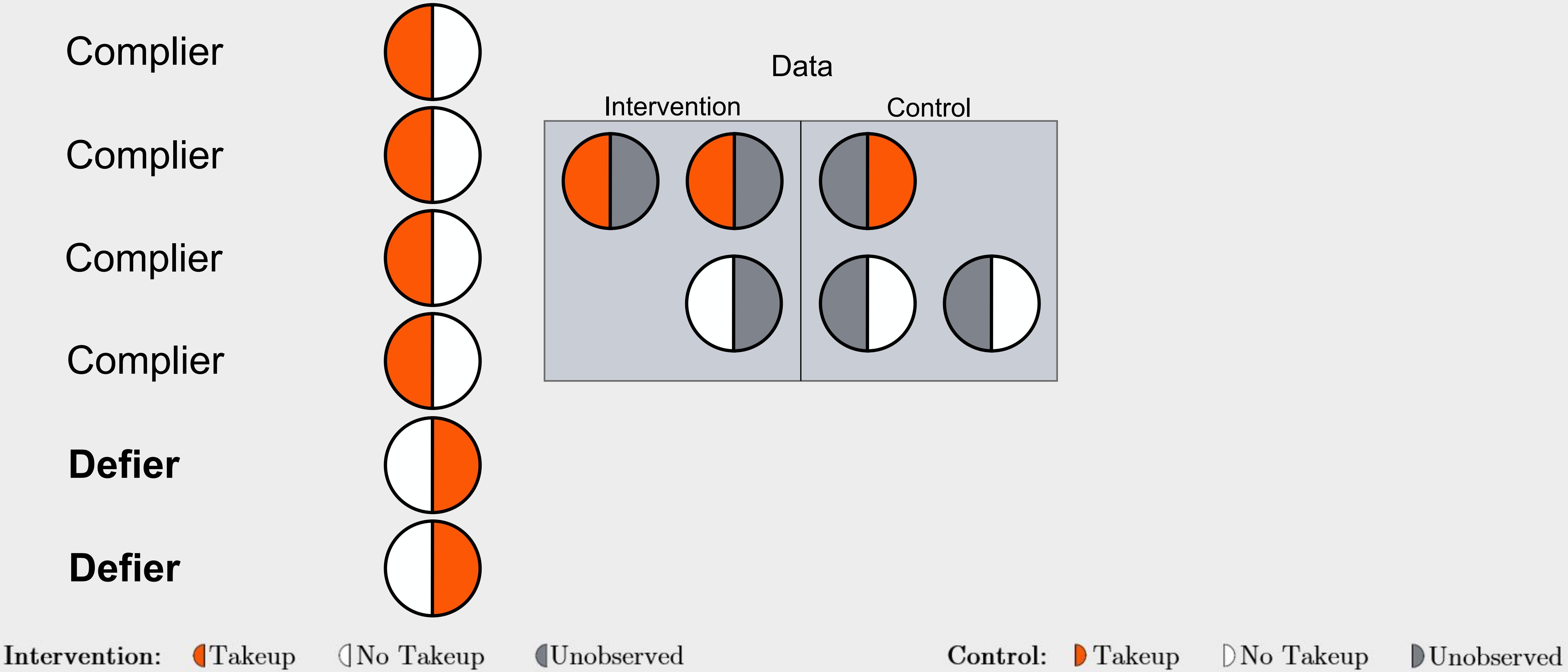
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

This is the optimal answer!

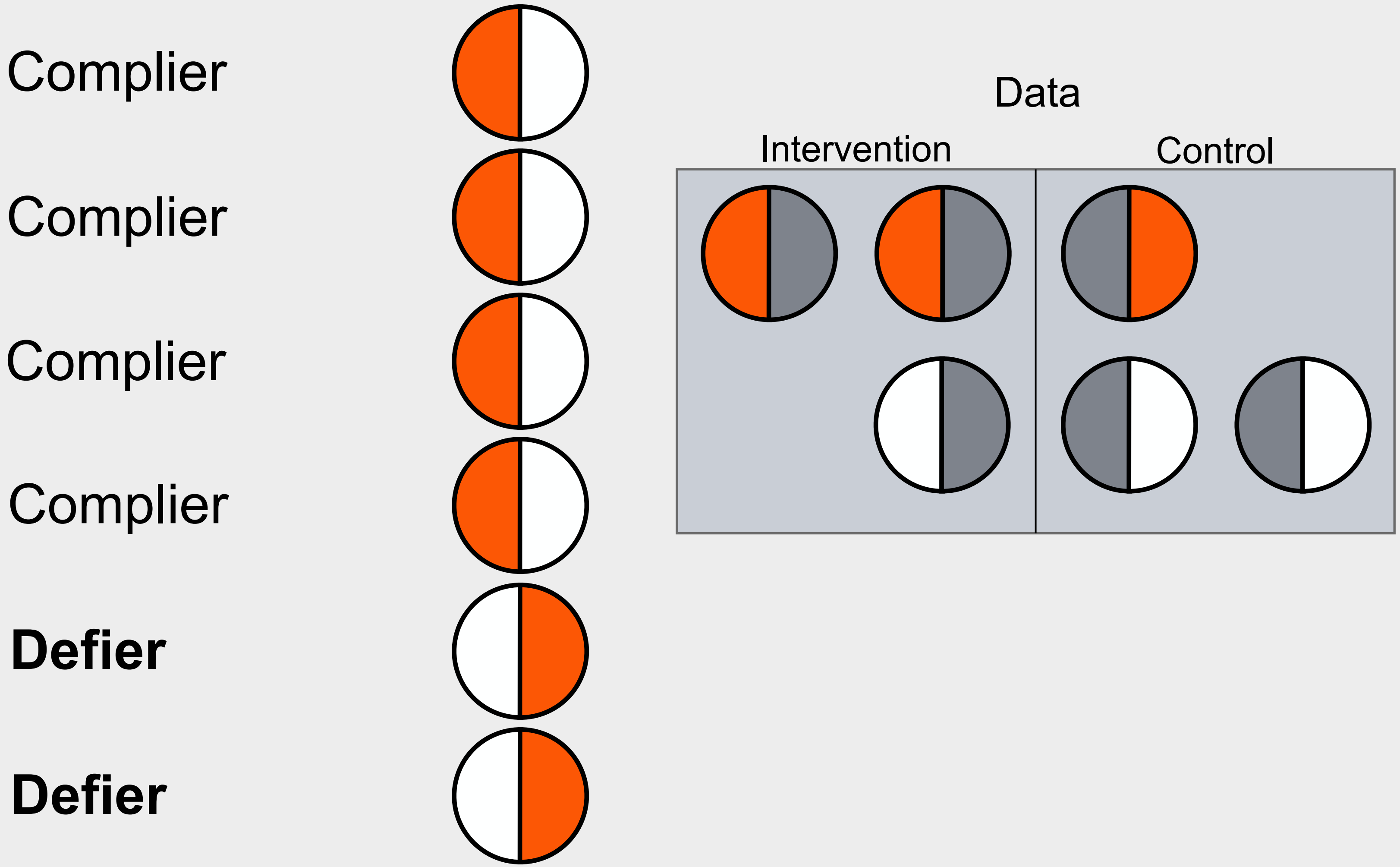


Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

Really? Why?

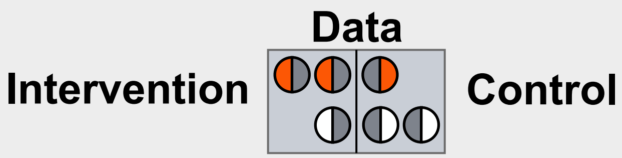

Complier
Complier
Complier
Complier
Defier
Defier
Data
Intervention
Control
Intervention: Takeup No Takeup Unobserved
Control: Takeup No Takeup Unobserved

We answer with our proposed design-based maximum likelihood decision rule.



Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

Here are the data again.

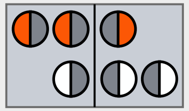


Intervention: Takeup No Takeup Unobserved

Control: Takeup No Takeup Unobserved

We can randomize three of the six people into intervention with 20 possible assignments.

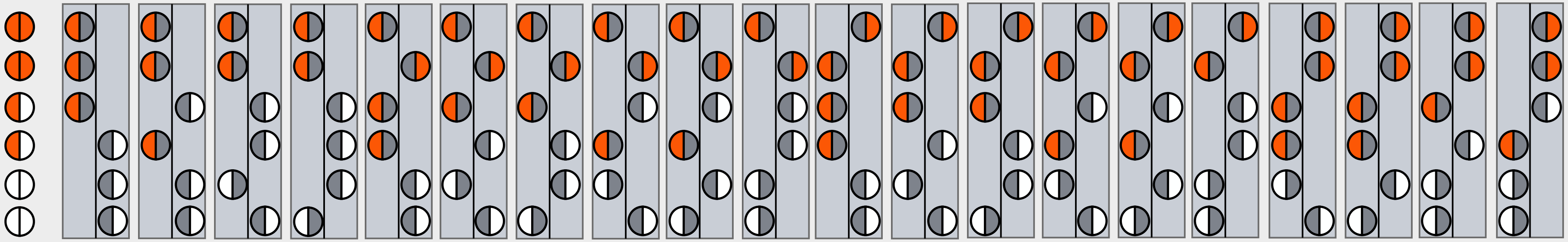
Data

Intervention  **Control**

Joint Distribution of Potential Outcomes

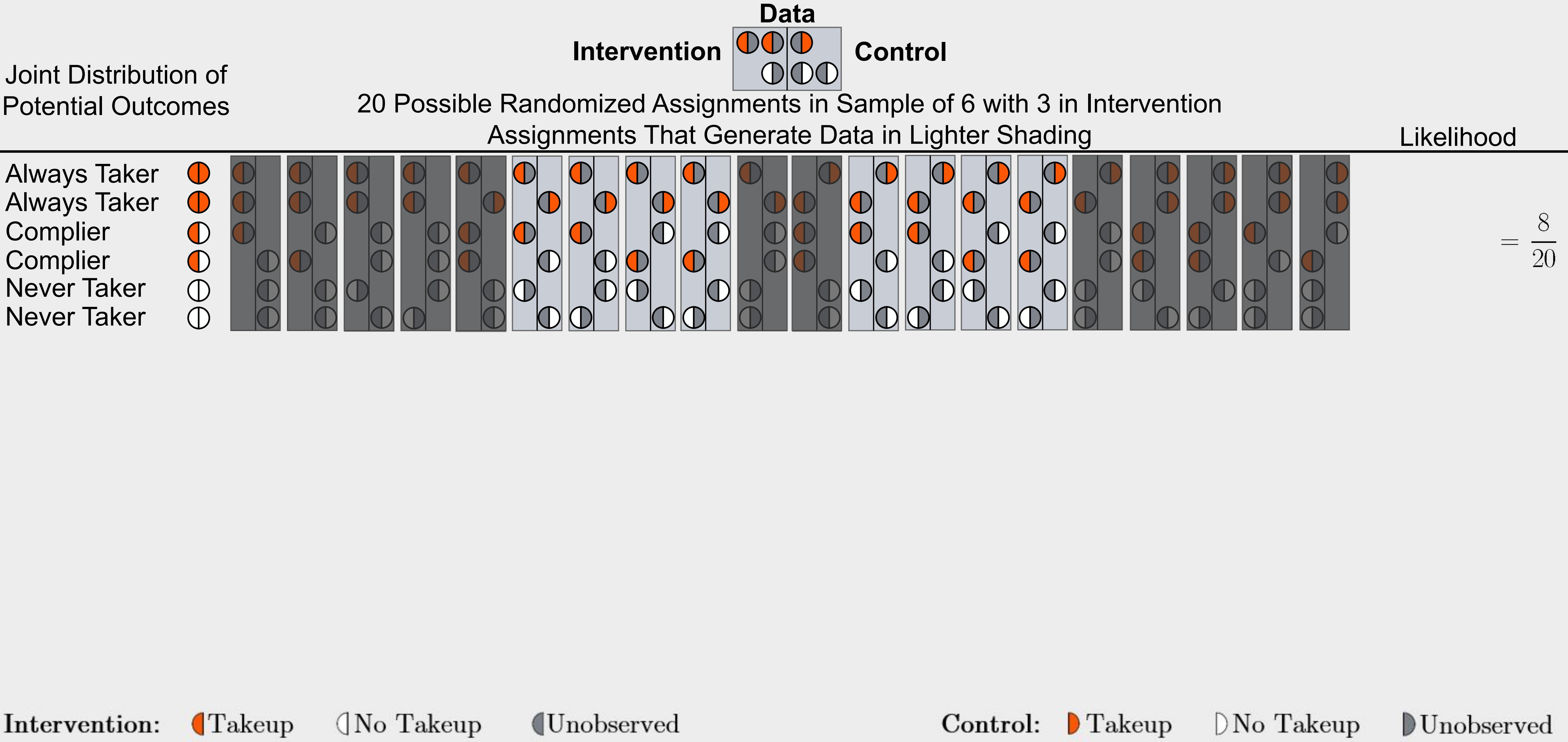
20 Possible Randomized Assignments in Sample of 6 with 3 in Intervention

Always Taker
Always Taker
Complier
Complier
Never Taker
Never Taker



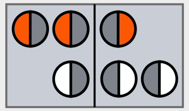
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

We can randomize three of the six people into intervention with 20 possible assignments. The eight assignments that generate the data are in lighter shading.



Assignments that generate the data have one always taker, one complier, and one never taker in intervention.

Data

Intervention  **Control**

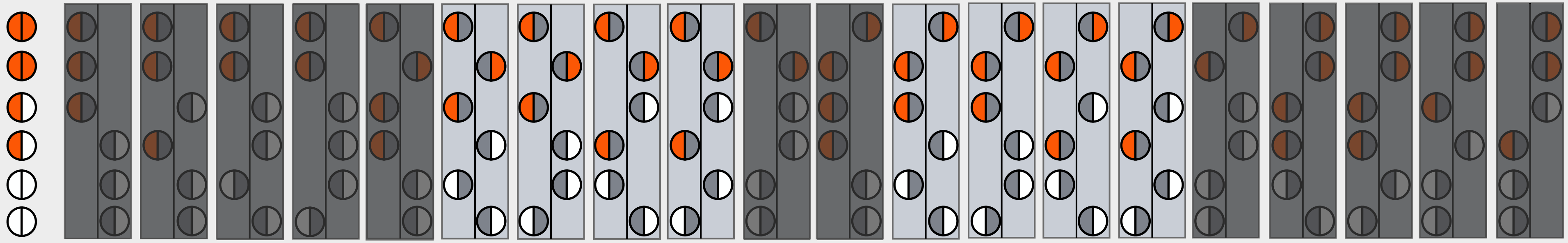
Joint Distribution of Potential Outcomes

20 Possible Randomized Assignments in Sample of 6 with 3 in Intervention

Assignments That Generate Data in Lighter Shading

Likelihood

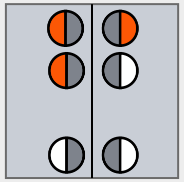
Always Taker
Always Taker
Complier
Complier
Never Taker
Never Taker



$$\frac{\binom{2}{1}\binom{2}{1}\binom{2}{1}}{\binom{6}{3}} = \frac{8}{20}$$

2 Always Takers, 2 Compliers, 2 Never Takers

1 Always Taker
1 Complier
0 Defiers
1 Never Taker

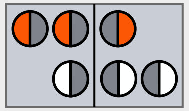


1 Always Taker
1 Complier
0 Defiers
1 Never Taker

Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

Assignments that generate the data have one always taker, one complier, and one never taker in intervention.

Data

Intervention  **Control**

Joint Distribution of Potential Outcomes

20 Possible Randomized Assignments in Sample of 6 with 3 in Intervention

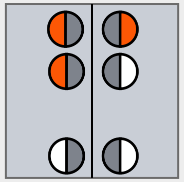
Assignments That Generate Data in Lighter Shading

Likelihood

Always Taker
Always Taker
Complier
Complier
Never Taker
Never Taker

2 Always Takers, 2 Compliers, 2 Never Takers

1 Always Taker
1 Complier
0 Defiers
1 Never Taker



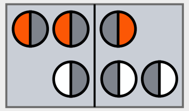
1 Always Taker
1 Complier
0 Defiers
1 Never Taker

$$\frac{\binom{2}{1}\binom{2}{1}\binom{2}{1}}{\binom{6}{3}} = \frac{8}{20}$$

Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

Here are the 20 possible randomized assignments for the answer with one defier.

Data

Intervention  **Control**

Joint Distribution of Potential Outcomes

20 Possible Randomized Assignments in Sample of 6 with 3 in Intervention

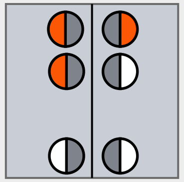
Assignments That Generate Data in Lighter Shading

Likelihood

Always Taker
Always Taker
Complier
Complier
Never Taker
Never Taker

2 Always Takers, 2 Compliers, 2 Never Takers

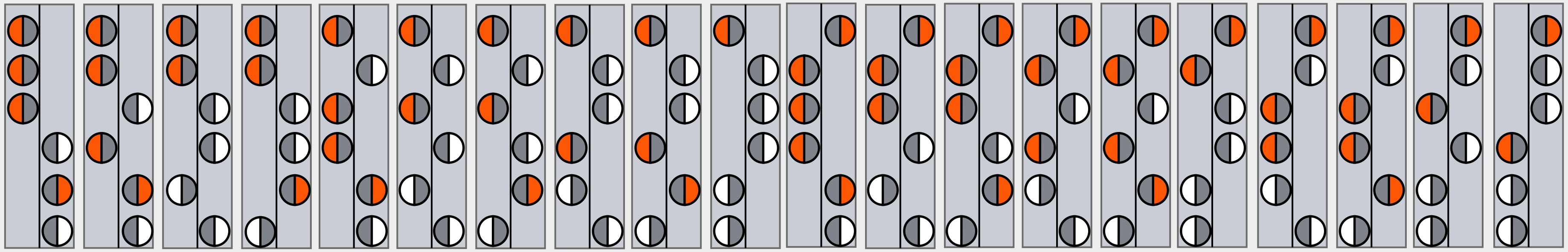
1 Always Taker
1 Complier
0 Defiers
1 Never Taker



1 Always Taker
1 Complier
0 Defiers
1 Never Taker

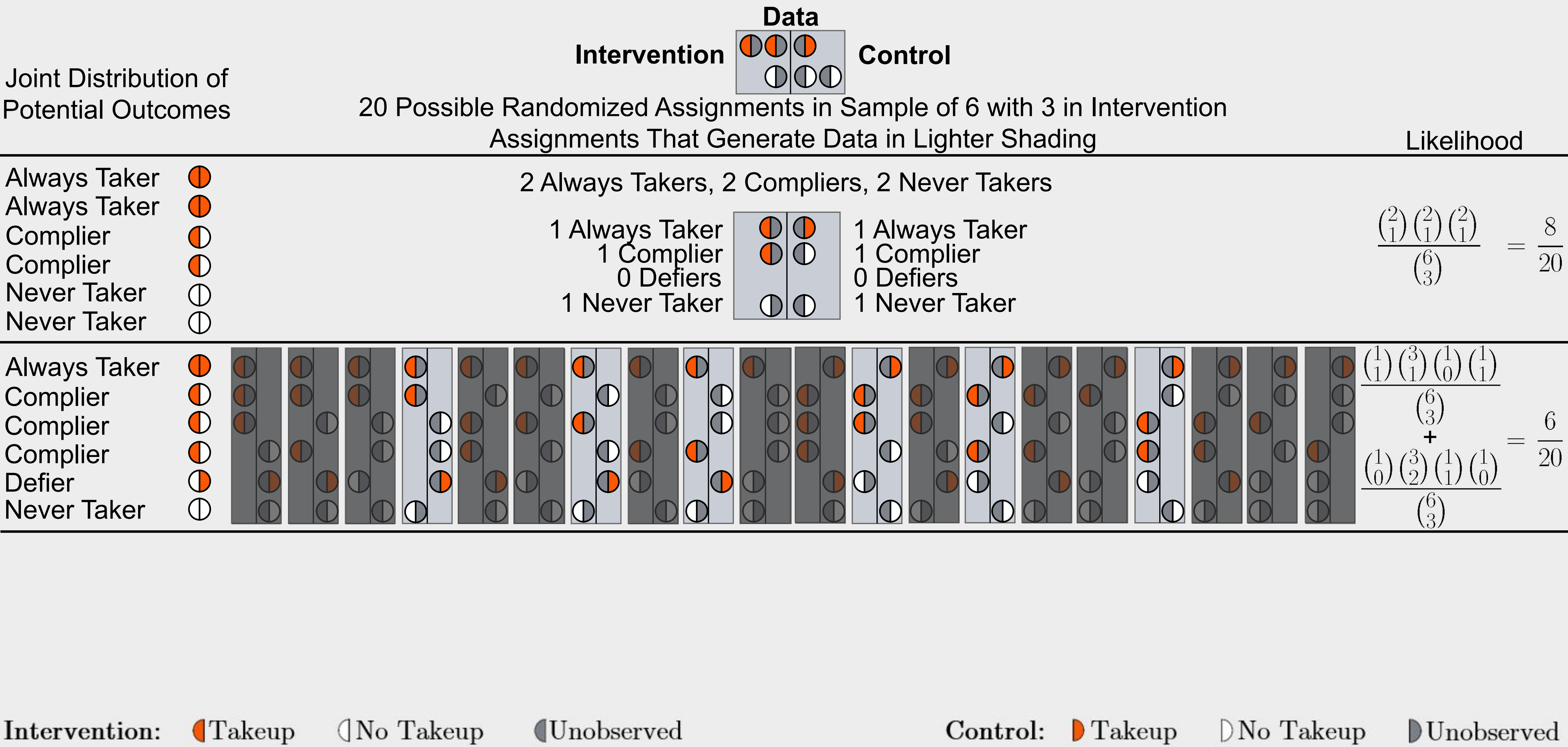
$$\frac{\binom{2}{1}\binom{2}{1}\binom{2}{1}}{\binom{6}{3}} = \frac{8}{20}$$

Always Taker
Complier
Complier
Complier
Defier
Never Taker



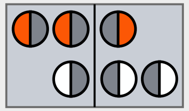
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

Only six assignments generate the data.



Only six assignments generate the data.

Data

Intervention  **Control**

Joint Distribution of Potential Outcomes

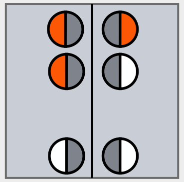
20 Possible Randomized Assignments in Sample of 6 with 3 in Intervention
Assignments That Generate Data in Lighter Shading

Likelihood

Always Taker
Always Taker
Complier
Complier
Never Taker
Never Taker

2 Always Takers, 2 Compliers, 2 Never Takers

1 Always Taker
1 Complier
0 Defiers
1 Never Taker



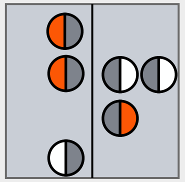
1 Always Taker
1 Complier
0 Defiers
1 Never Taker

$$\frac{\binom{2}{1}\binom{2}{1}\binom{2}{1}}{\binom{6}{3}} = \frac{8}{20}$$

Always Taker
Complier
Complier
Complier
Defier
Never Taker

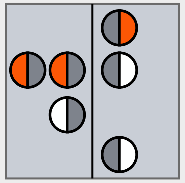
1 Always Taker, 3 Compliers, 1 Defier, 1 Never Taker

1 Always Taker
1 Complier
0 Defiers
1 Never Taker



0 Always Takers
2 Compliers
1 Defier
0 Never Takers

0 Always Takers
2 Compliers
1 Defier
0 Never Taker



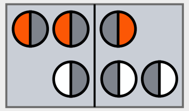
1 Always Taker
1 Complier
0 Defiers
1 Never Taker

$$\frac{\binom{1}{1}\binom{3}{1}\binom{1}{0}\binom{1}{1}}{\binom{6}{3}} + \frac{\binom{1}{0}\binom{3}{2}\binom{1}{1}\binom{1}{0}}{\binom{6}{3}} = \frac{6}{20}$$

Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

Now consider the solution with two defiers. How many assignments generate the data?

Data

Intervention  **Control**

Joint Distribution of Potential Outcomes

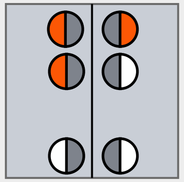
20 Possible Randomized Assignments in Sample of 6 with 3 in Intervention
Assignments That Generate Data in Lighter Shading

Likelihood

Always Taker
Always Taker
Complier
Complier
Never Taker
Never Taker

2 Always Takers, 2 Compliers, 2 Never Takers

1 Always Taker
1 Complier
0 Defiers
1 Never Taker



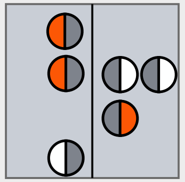
1 Always Taker
1 Complier
0 Defiers
1 Never Taker

$$\frac{\binom{2}{1}\binom{2}{1}\binom{2}{1}}{\binom{6}{3}} = \frac{8}{20}$$

Always Taker
Complier
Complier
Complier
Defier
Never Taker

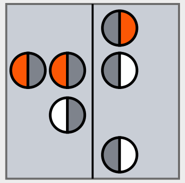
1 Always Taker, 3 Compliers, 1 Defier, 1 Never Taker

1 Always Taker
1 Complier
0 Defiers
1 Never Taker



0 Always Takers
2 Compliers
1 Defier
0 Never Takers

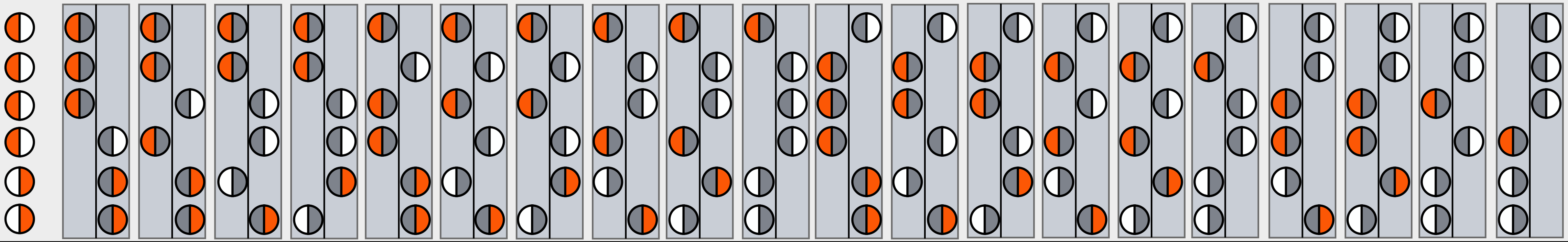
0 Always Takers
2 Compliers
1 Defier
0 Never Taker



1 Always Taker
1 Complier
0 Defiers
1 Never Taker

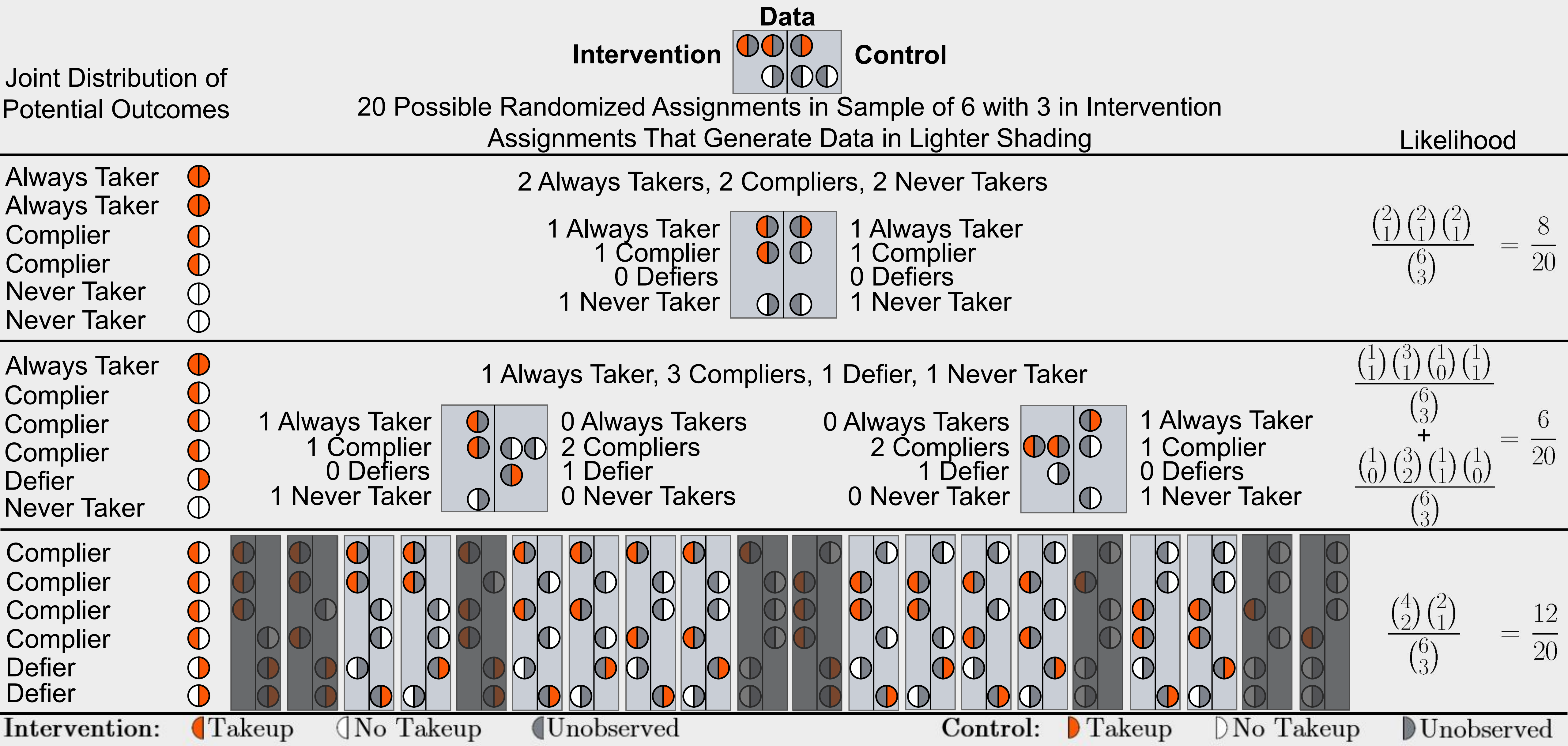
$$\frac{\binom{1}{1}\binom{3}{1}\binom{1}{0}\binom{1}{1}}{\binom{6}{3}} + \frac{\binom{1}{0}\binom{3}{2}\binom{1}{1}\binom{1}{0}}{\binom{6}{3}} = \frac{6}{20}$$

Complier
Complier
Complier
Complier
Defier
Defier



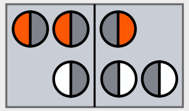
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

12 of the 20 possible assignments generate the data!



The answer with two defiers has the maximum likelihood!
We decide that the intervention backfires for two people in the sample.

Data

Intervention  **Control**

Joint Distribution of Potential Outcomes

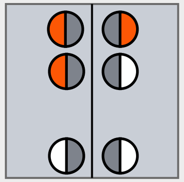
20 Possible Randomized Assignments in Sample of 6 with 3 in Intervention
Assignments That Generate Data in Lighter Shading

Likelihood

Always Taker
Always Taker
Complier
Complier
Never Taker
Never Taker

2 Always Takers, 2 Compliers, 2 Never Takers

1 Always Taker
1 Complier
0 Defiers
1 Never Taker



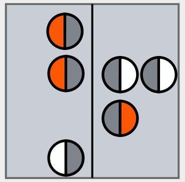
1 Always Taker
1 Complier
0 Defiers
1 Never Taker

$$\frac{\binom{2}{1}\binom{2}{1}\binom{2}{1}}{\binom{6}{3}} = \frac{8}{20}$$

Always Taker
Complier
Complier
Complier
Defier
Never Taker

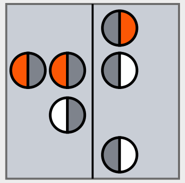
1 Always Taker, 3 Compliers, 1 Defier, 1 Never Taker

1 Always Taker
1 Complier
0 Defiers
1 Never Taker



0 Always Takers
2 Compliers
1 Defier
0 Never Takers

0 Always Takers
2 Compliers
1 Defier
0 Never Taker



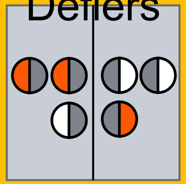
1 Always Taker
1 Complier
0 Defiers
1 Never Taker

$$\frac{\binom{1}{1}\binom{3}{1}\binom{1}{0}\binom{1}{1}}{\binom{6}{3}} + \frac{\binom{1}{0}\binom{3}{2}\binom{1}{1}\binom{1}{0}}{\binom{6}{3}} = \frac{6}{20}$$

Complier
Complier
Complier
Complier
Defier
Defier

4 Compliers, 2 Defiers

0 Always Takers
2 Compliers
1 Defier
0 Never Taker



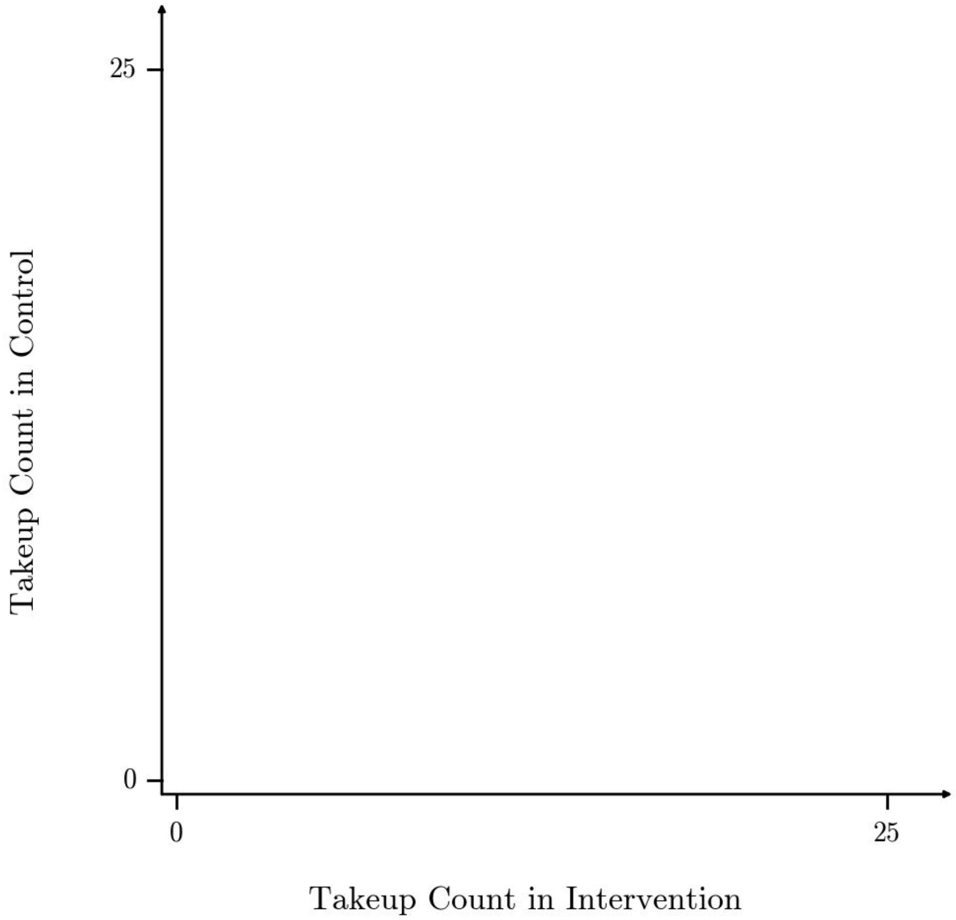
0 Always Takers
2 Compliers
1 Defier
0 Never Takers

$$\frac{\binom{4}{2}\binom{2}{1}}{\binom{6}{3}} = \frac{12}{20}$$

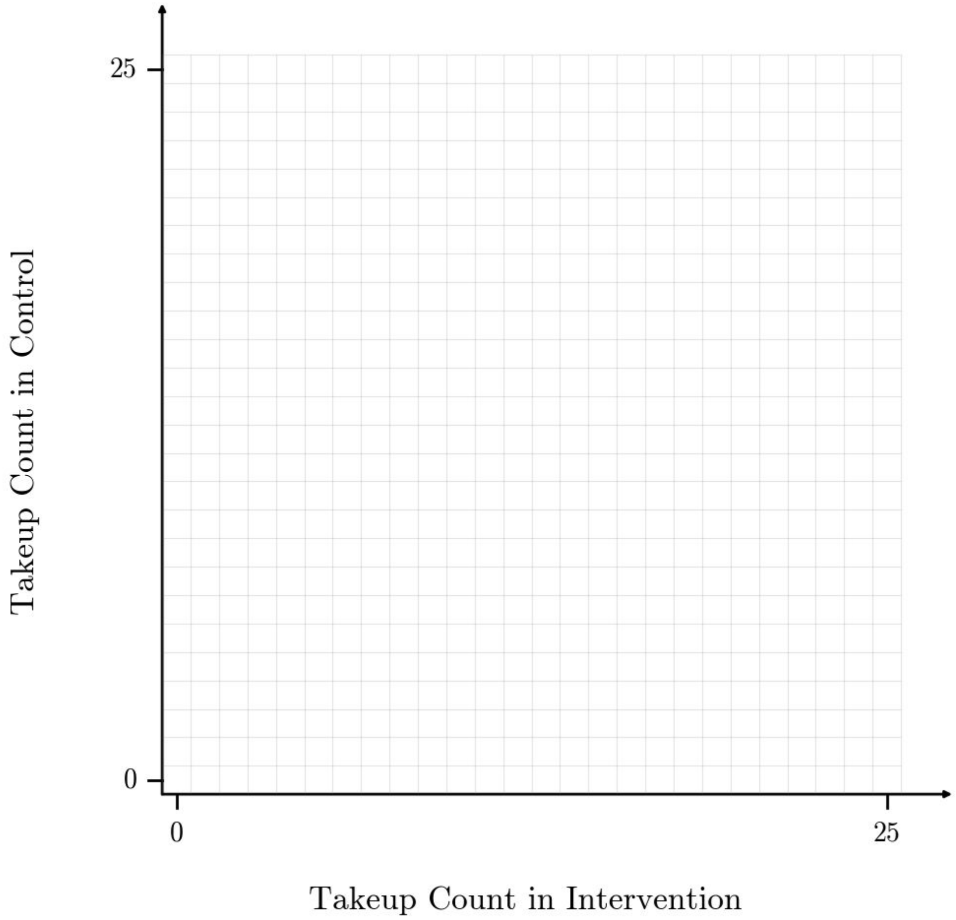
Intervention: Takeup No Takeup Unobserved **Control:** Takeup No Takeup Unobserved

What can we say beyond the average effect in all possible results with 50 people, 25 in intervention?

These axes show the takeup count in intervention and the takeup count in control.

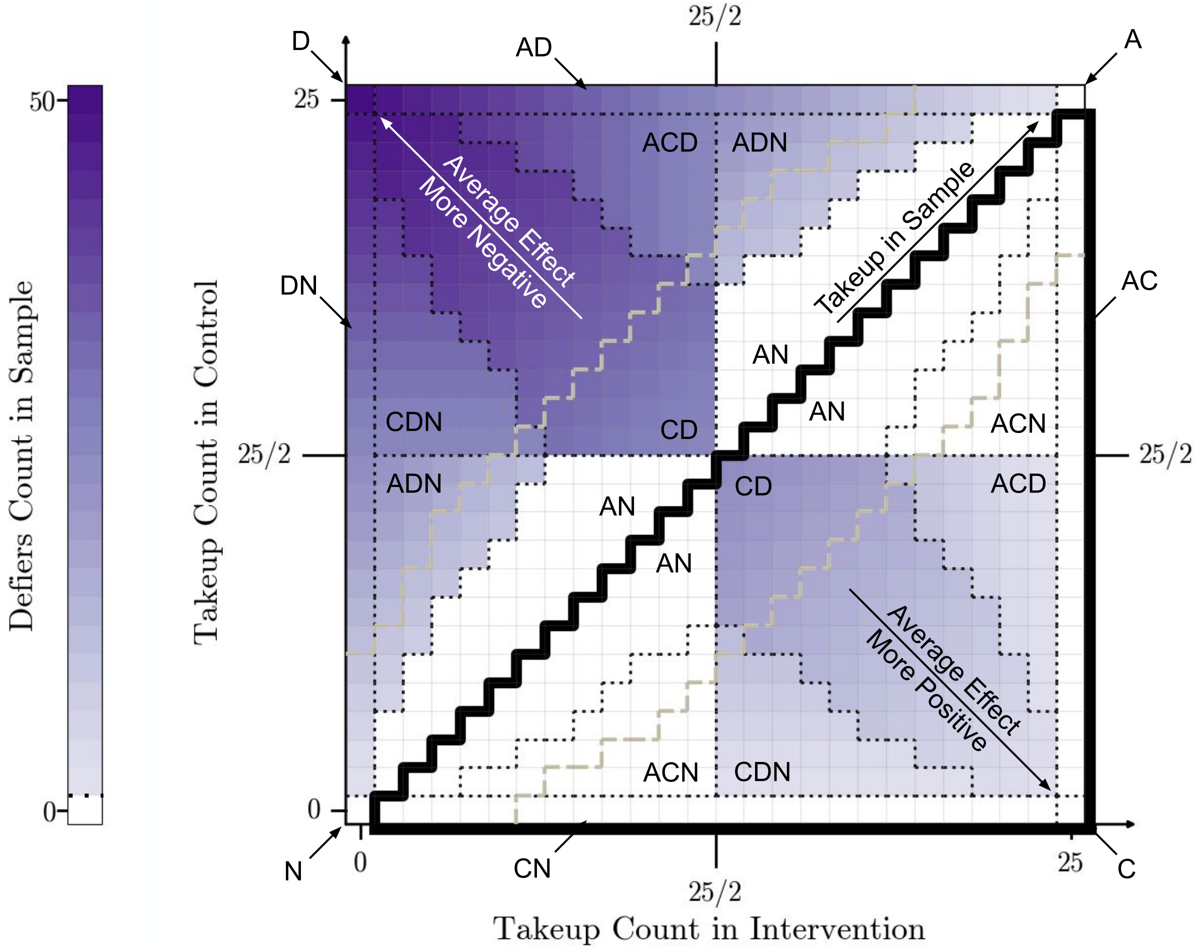


There are the 676 (=26x26) possible realizations of the data in a sample of 50 people.



We can reveal evidence beyond the average effect!

Region with Positive Average Effect	Regions with Same Types Present in MLE	Region where Fisher's Exact Test Fails to Reject Null at 5% Level



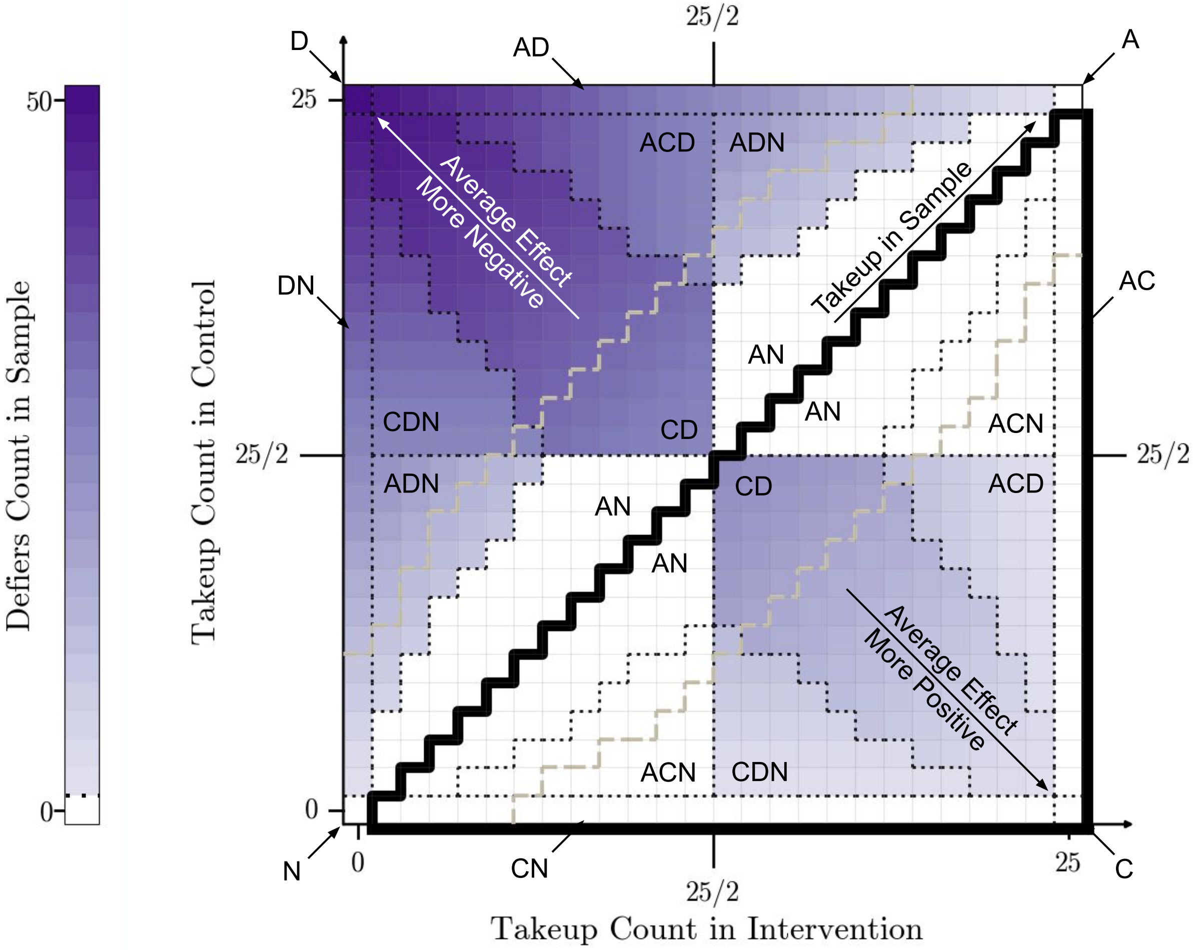
Types: A: Always Takers C: Compliers D: Defiers N: Never Takers

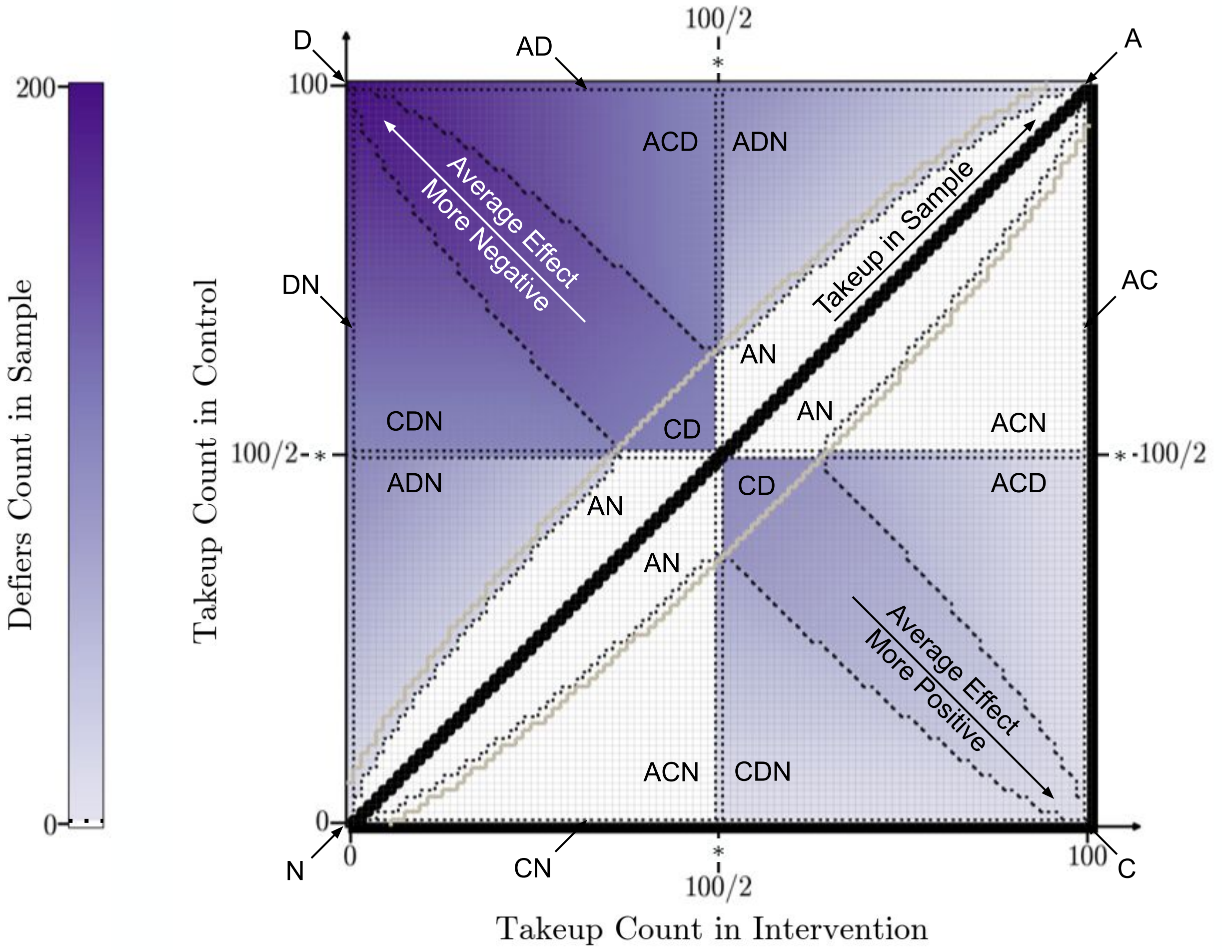
What happens to the patterns if the sample size quadruples to 200?

Region with Positive Average Effect

Regions with Same Types Present in MLE

Region where Fisher's Exact Test Fails to Reject Null at 5% Level





* Two-way ties in the MLE occur in cells bisected by the 100/2 lines, between MLEs in the cells on either side of the line, only when takeup count in intervention or control is not zero or full.

Types: A: Always Takers C: Compliers D: Defiers N: Never Takers

In this published experiment that offered a payment to pregnant smokers, we decide that the intervention does not backfire for anyone in the sample.

	Tappin et al. (2015)	Johnson and Goldstein (2003)
Context and Design		
Takeup	Quit Smoking	
Intervention	Payment	
Control	Usual Care	
Sample	Pregnant Smokers	
Design	Completely Randomized	
Standard Statistics		
Average Effect	$69/306 - 26/306 = 14\%$	
[95% Confidence Interval]	$[8\%, 20\%]$	
Fisher's Exact Test p-value	< 0.001	
Intervention Takeup Rate	$69/306 = 23\%$	
Control Takeup Rate	$26/306 = 8\%$	
Sample Size	612	
Proposed Design-Based Maximum Likelihood Estimates		
Always Takers	52/612=8%	
[95% Smallest Credible Set]	[0/612, 72/612]=[0%, 12%]	
Compliers	86/612=14%	
[95% Smallest Credible Set]	[44/612, 164/612]=[7%, 27%]	
Defiers	0/612=0%	
[95% Smallest Credible Set]	[0/612, 71/612]=[0%, 12%]	
Never Takers	474/612=77%	
[95% Smallest Credible Set]	[388/612, 498/612]=[63%, 81%]	

Percentages are rounded to whole numbers. We construct a 95% smallest credible set on the number of each type in the sample by reporting the collection of values for each type occurring within the 95% smallest credible set on the joint distribution of potential outcomes. Because the numbers of each type must sum to the sample size, the 95% smallest credible sets for each of the four types are dependent.

But in this well-known experiment that offered an opt-out nudge to become an organ donor, we decide that the intervention does backfire for 18% of the sample.

	Tappin et al. (2015)	Johnson and Goldstein (2003)
Context and Design		
Takeup	Quit Smoking	Become Organ Donor
Intervention	Payment	Opt-out
Control	Usual Care	Opt-in
Sample	Pregnant Smokers	Online Respondents
Design	Completely Randomized	Bernoulli Randomized
Standard Statistics		
Average Effect	$69/306 - 26/306 = 14\%$	$50/61 - 23/54 = 39\%$
[95% Confidence Interval]	$[8\%, 20\%]$	$[23\%, 56\%]$
Fisher's Exact Test p-value	< 0.001	< 0.001
Intervention Takeup Rate	$69/306 = 23\%$	$50/61 = 82\%$
Control Takeup Rate	$26/306 = 8\%$	$23/54 = 43\%$
Sample Size	612	115
Proposed Design-Based Maximum Likelihood Estimates		
Always Takers	52/612=8%	28/115=24%
[95% Smallest Credible Set]	[0/612, 72/612]=[0%, 12%]	[0/115, 63/115]=[0%, 55%]
Compliers	86/612=14%	66/115=57%
[95% Smallest Credible Set]	[44/612, 164/612]=[7%, 27%]	[23/115, 81/115]=[20%, 70%]
Defiers	0/612=0%	21/115=18%
[95% Smallest Credible Set]	[0/612, 71/612]=[0%, 12%]	[0/115, 34/115]=[0%, 30%]
Never Takers	474/612=77%	0/115=0%
[95% Smallest Credible Set]	[388/612, 498/612]=[63%, 81%]	[0/115, 32/115]=[0%, 28%]

Percentages are rounded to whole numbers. We construct a 95% smallest credible set on the number of each type in the sample by reporting the collection of values for each type occurring within the 95% smallest credible set on the joint distribution of potential outcomes. Because the numbers of each type must sum to the sample size, the 95% smallest credible sets for each of the four types are dependent.

Fisher 1935:

The **design** of Charles Darwin's experiments was "greatly superior" to "statistical methods available at the time."

Athey and Imbens 2017:

"We recommend using statistical methods that are directly justified by randomization, in contrast to the more traditional sampling-based approach that is commonly used in econometrics."

Counting Defiers: A Design-Based Model of an Experiment Can Reveal Evidence Beyond the Average Effect

We estimate the numbers of always takers, compliers, defiers, and never takers in the sample

- The sampling-based literature considers the shares of these types in a superpopulation.
- It is well-known that the sampling-based likelihood does not vary with the share of defiers within the Boole 1854, Hoeffding 1940, and Fréchet 1957 bounds.
- Sampling-based literature considers bounds on shares of defiers in the population.
    - Balke and Pearl 1997, Heckman, Smith and Clements 1997, Manski 1997, Tian and Pearl 2000, Zhang and Rubin 2003, Imbens and Manski 2004, Fan and Park 2010, Mullahy 2018, Ding and Miratrix 2019. Li and Pearl 2019, Bai, Huang, Shaikh, and Vytlacil 2024, Semenova 2024.

Counting Defiers: A Design-Based Model of an Experiment Can Reveal Evidence Beyond the Average Effect

We consider a primitive estimand: always takers, compliers, defiers, and never takers in the sample

1. Using a causal model with parametric structure from the randomization design
    - Neyman 1923, Welch 1937, Kempthorne 1952, Copas 1973, Rubin 1974, 1977, Greenland and Robins 1986, Holland 1986, and others develop the causal model.
    - Copas 1973 uses the design-based likelihood to compare hypothesis tests.
    - Li and Ding 2016 use parametric structure to construct exact confidence intervals on the average effect.

Counting Defiers: A Design-Based Model of an Experiment Can Reveal Evidence Beyond the Average Effect

We consider a primitive estimand: always takers, compliers, defiers, and never takers in the sample

1. Using a causal model with parametric structure from the randomization design
2. And a maximum likelihood decision rule that can harness weak evidence
    - Fisher 1935 exact test discards this weak evidence.
        - Tetenov 2012 discusses how implied rules from all tests treat Type I and II errors asymmetrically.
    - Our rule varies with the data.
        - Ferguson 1967: our rule is “reasonable” because it is “better than just guessing.”
        - Minimax and minimax regret rules do not necessarily vary with data.
            - Schalg 2003, Manski 2004, Hirano and Porter 2009, and Stoye 2009.
    - Our rule is Bayes optimal.
        - Our rule is admissible (Ferguson 1967).
        - Our rule cannot be bested in a betting framework (Freedman and Purves 1969).
        - We can quantify the gains of our rule over a Fréchet rule and a monotonicity rule.
        - We can construct credible sets as an alternative to sampling-based inference on Fréchet bounds (Manski, Sandefur, McLanahan, and Powers 1992, Horowitz and Manski 2000, Tamer, Chernozhukov, and Hong 2004) and defiers within the Fréchet bounds (Imbens and Manski 2004)
    - Our rule is optimal by the principle of maximum entropy (Jaynes 1957a,b).
        - Golan 2002 uses entropy to abstract away from parametric structure, which we embrace.

Counting Defiers: A Design-Based Model of an Experiment Can Reveal Evidence Beyond the Average Effect

We consider a primitive estimand: always takers, compliers, defiers, and never takers in the sample

1. Using a causal model with parametric structure from the randomization design
2. And a maximum likelihood decision rule that can harness weak evidence
    - Our rule contributes to the integration of statistical decision theory into econometrics,
        - Manski 2004, Dehejia 2005, Manski 2007, Hirano 2008, Hirano and Porter 2009, Stoye 2012, Kitagawa and Tetenov 2018, Manski 2018, 2019, Manski and Tetenov 2021, Fernandez, Montiel Olea, Qiu, Stoye, and Tinda 2024
    - Particularly in finite sample settings.
        - Canner 1970, Manski and Tetenov 2007, Schlag 2007, Stoye 2007, 2009, Tetenov 2012.

Counting Defiers: A Design-Based Model of an Experiment Can Reveal Evidence Beyond the Average Effect

We consider a primitive estimand: always takers, compliers, defiers, and never takers in the sample

1. Using a causal model with parametric structure from the randomization design
2. And a maximum likelihood decision rule that can harness weak evidence
3. To support a monotonicity assumption or a specific alternative.
   - Previous evidence against monotonicity requires more data than a binary intervention and outcome.
     - Machine learning requires covariates.
       - For example: Wager and Athey 2018, Semenova 2024.
     - Analysis of side effects in medicine requires secondary outcomes.
       - For example, Barnard et al. 2001.
     - Other approaches require data on a second stage outcome and a two-stage model.
       - Specification tests reveal evidence against model and/or monotonicity in the first stage.
         - Imbens and Rubin 1997, Richardson and Robins 2010, Huber and Mellace 2012, 2015, Kitagawa 2015, Mourifié and Wan 2017, Machado, Shaikh, and Vytlacil 2019.
       - Marginal treatment effect models can reveal evidence against monotonicity in the second stage under instrumental variable assumptions plus ancillary assumptions.
         - Bjorklund and Moffit 1987, Heckman and Vytlacil 1999, Kowalski 2023a,b.
       - Chan, Gentzkow, and Yu (2022) require a multi-valued instrument and a second stage outcome for at least one value of the first stage outcome.

Counting Defiers: A Design-Based Model of an Experiment Can Reveal Evidence Beyond the Average Effect

We consider a primitive estimand: always takers, compliers, defiers, and never takers in the sample

1. Using a causal model with parametric structure from the randomization design
2. And a maximum likelihood decision rule that can harness weak evidence
3. To support a monotonicity assumption or a specific alternative.

Our goal is to improve the average effect of future interventions by targeting away from defiers.

Our work has implications for methodological and applied work.

- Derive design-based likelihoods for other designs: two stage models, matched pairs, stratified, and permuted block to obtain exact test statistics and confidence intervals
- Improve computation.
- Consider other utility functions
  - Outcome quantiles (Guggenberger, Mehta, and Pavlov 2024)
  - Welfare (Cui and Han 2023)
  - Asymmetric counterfactual utilities (Ben-Michael, Imai, Jiang 2024, Christy and Kowalski 2024, Gelman and Mikhaeil 2025)
- Leverage our visualizations, which we see as a secondary contribution.
- Revisit results on optimal experimental design that focus on the average effect (Bai 2022).
- Report randomization design and sample counts in applied work.

Counting Defiers: A Design-Based Model of an Experiment Can Reveal Evidence Beyond the Average Effect

Neil Christy and Amanda Kowalski
University of Michigan

Thanks to Jann Spiess for extensive regular feedback and to Charles Manski, Toru Kitagawa, Aleksey Tetenov, and Donald Rubin for encouraging us to use statistical decision theory and teaching us about it. Thanks also to Guido Imbens.

Paper on arXiv: https://arxiv.org/abs/2412.16352