Data and Causation

EC 350: Labor Economics

Kyle Raze

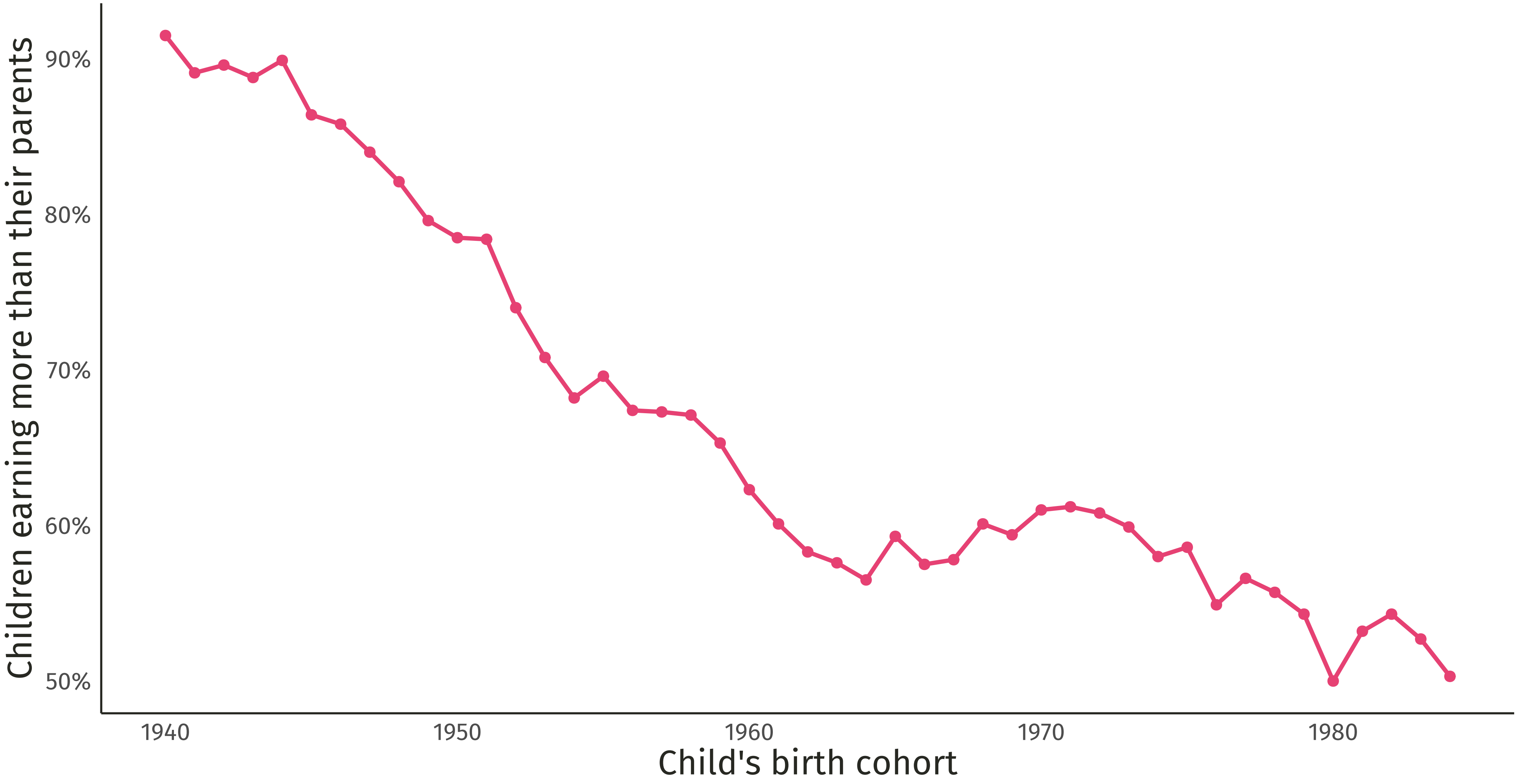
Winter 2022

Data and Causation

1. The rise of empirical evidence
2. Making *other-things-equal* comparisons
3. What do "data" look like?
4. Causal identification
    - Average treatment effects
    - Selection bias
5. Randomized control trials
6. *Thinking Fast and Slow*, Chicago edition

The rise of empirical evidence

The fading *American dream*



Source: Raj Chetty et al. (2017), The fading American dream: Trends in absolute income mobility since 1940, *Science*.

Why is the *American dream* fading?

Policy Question: Why is a child's chance of climbing the income ladder decreasing in the United States?

- What can we do to reverse this trend?

Difficult to answer with historical data on macroeconomic trends.

- **That other things change over time makes it difficult** to separately identify the roles of alternative explanations

Theoretical social science

Historically, the social sciences had **limited data** to study policy questions.

The result? Social sciences were theoretical fields

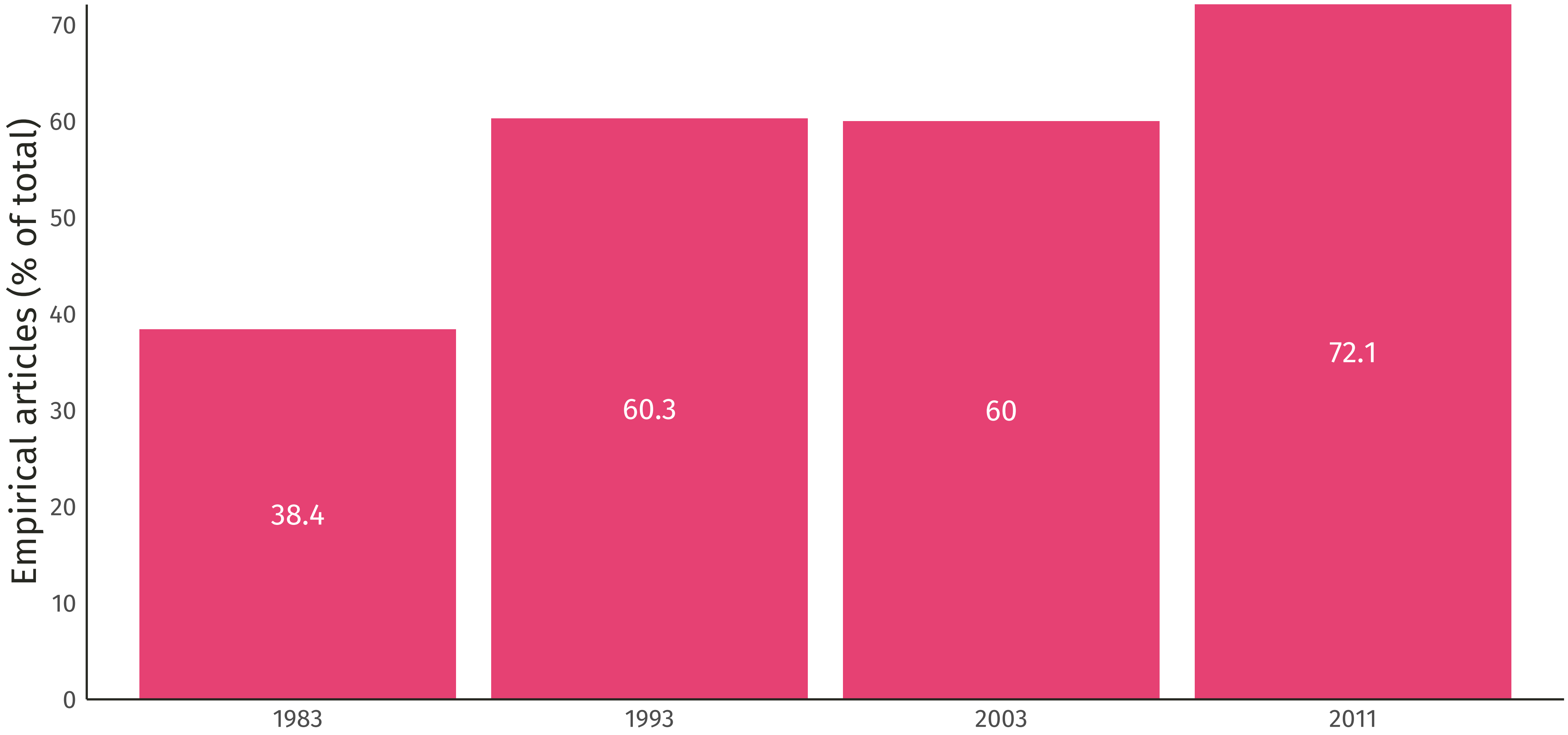
- Some researchers developed mathematical models
- Some developed qualitative theories
- Both used their theories to make policy recommendations (*e.g.*, to improve upward mobility)

The problem? Without data, even falsifiable theories are never tested!

- Five researchers could have five different answers to the same question
- Can lead to a politicization of questions that, in principle, have scientific answers (*e.g.*, do minimum wage laws reduce employment?)

Economics is becoming more data-driven

Empirical articles in the top three economics journals over time



Source: Daniel S. Hamermesh (2013), Six Decades of Top Economics Publishing: Who and How? *Journal of Economic Literature*.

The rise of empirical evidence

Today, the social sciences are increasingly **empirical** thanks to the growing availability of data and computational power.

- Gives us **the ability to test** existing theories
- Gives us **the ability to refine** theory to (i) better explain decision making and (ii) better fit real-world data

The social sciences have caught up to the natural sciences in terms of scientific rigor, arguably surpassing the natural sciences in sophistication.

- Given the complexity of human decision making, inability to experiment in controlled environments, *etc.*?

Making *other-things-equal* comparisons

Other-things-equal comparisons

The policy? In 2017, the University of Oregon started requiring first-year students to live on campus.

The rationale? First-year students who live on campus outperform those who live off campus.

- Average 2nd-year retention rate *5 percentage points higher*
- *80 percent more likely to graduate* in four years
- GPA *0.13 points higher*

Q: Do these comparisons suggest that the policy will improve student outcomes?

Q: Do they describe the effect of living on campus?

Q: Do they describe *something else*?

Other-things-equal comparisons

Healthy skepticism should leave us questioning the UO's interpretation.

- The **decision** to live on campus is likely related to family wealth and interest in school.
- Family wealth and interest in school are also related to academic achievement.

The difference in outcomes between those on and off campus **does not offer an *other-things-equal* comparison**.

- Without further evidence, one should not attribute the difference in outcomes to living on campus.
  - Not without considering those things that both (i) correlate with living on campus (*e.g.*, family wealth) and (ii) correlate with outcomes (*e.g.*, graduation)

Other-things-equal comparisons

Statistical comparisons can only identify causal relationships between variables **when all other factors are "held constant."**

- *Causal* relationship = How a change in one variable *induces* a change in another

Economists have developed a *comparative advantage*[†] in understanding where **other-things-equal** comparisons can (and cannot) be made.

- Anyone can retort "*correlation doesn't imply causation!*"
- Understanding why it doesn't? The conditions under which it actually does imply causality?
  - Difficult, but necessary for learning from data!

[†] *Comparative advantage* = Ability of an individual or group to perform an activity at lower cost relative to another individual or group.

What do "data" look like?

What do "data" look like?

Sample of US workers (1976)

	Wage	Education	Female?
1	3.1	11	1
2	3.24	12	1
3	3	11	0
4	6	8	0
5	5.3	12	0

Showing 1 to 5 of 526 entries

Previous 1 2 3 4 5 … 106 Next

Rows represent **observations**.

Columns represent **variables**.

Each value is associated with an **observation** and a **variable**.

Causal identification

Causal identification

The objective

Identify the effect of a **treatment** on an **outcome**.

The ideal comparison

Ideally, we could calculate the **treatment effect** *for each individual* as

$$Y_{1,i} - Y_{0,i}$$

- $Y_{1,i}$ is the outcome for person i when i receives the treatment
- $Y_{0,i}$ is the outcome for person i when i does not receive the treatment
- Known as **potential outcomes**

Causal identification

The **ideal data** for 10 people

```
#>     i treat Y_1i Y_0i effect_i
#> 1   1     1 5.01 4.56     0.45
#> 2   2     1 8.85 4.53     4.32
#> 3   3     1 6.31 4.67     1.64
#> 4   4     1 5.97 4.79     1.18
#> 5   5     1 7.61 6.34     1.27
#> 6   6     0 7.63 4.15     3.48
#> 7   7     0 4.75 0.56     4.19
#> 8   8     0 5.77 3.52     2.25
#> 9   9     0 7.47 4.49     2.98
#> 10 10     0 7.79 1.40     6.39
```

We could calculate the treatment effect for each individual i,

$$\tau_i = Y_{1,i} - Y_{0,i} \ ,$$

and we would be inclined to think of it as the causal effect.

The mean of these individual treatment effects = 2.82

- We call this the **average treatment effect** (ATE)

Causal identification

The fundamental problem of causal inference

While the ideal comparison is

$$\tau_i = Y_{1,i} - Y_{0,i} \ ,$$

this comparison is fundamentally challenged!

- If we observe Y_1 for i, then we cannot observe Y_0 for i
- If we observe Y_0 for i, then we cannot observe Y_1 for i
- We only observe what *actually* happened—we cannot observe the **counterfactual**

The implication? ALL causal inference is **by assumption!**

Causal identification

The data we *actually* see for these 10 people?

```
#>     i treat Y_1i Y_0i
#> 1   1     1 5.01   NA
#> 2   2     1 8.85   NA
#> 3   3     1 6.31   NA
#> 4   4     1 5.97   NA
#> 5   5     1 7.61   NA
#> 6   6     0   NA 4.15
#> 7   7     0   NA 0.56
#> 8   8     0   NA 3.52
#> 9   9     0   NA 4.49
#> 10 10     0   NA 1.40
```

We only observe Y_1 for $i \in \{1, \ldots, 5\}$

We only observe Y_0 for $i \in \{6, \ldots, 10\}$

We do not observe both $Y_{1,i}$ and $Y_{0,i}$ for anyone

Q: How can we estimate the average treatment effect when we cannot observe individual treatment effects?

Causal identification

Can we **compare the mean outcomes** of each group?

- Take the average of Y_1 for those who received the treatment (*i.e.*, the treatment-group mean)
- Take the average of Y_0 for those who didn't receive the treatment (*i.e.*, the control-group mean)

Q: Does treatment-group mean − control-group mean isolate the causal effect of the treatment?

Causal identification

```
#>     i treat Y_1i Y_0i
#> 1   1     1 5.01   NA
#> 2   2     1 8.85   NA
#> 3   3     1 6.31   NA
#> 4   4     1 5.97   NA
#> 5   5     1 7.61   NA
#> 6   6     0   NA 4.15
#> 7   7     0   NA 0.56
#> 8   8     0   NA 3.52
#> 9   9     0   NA 4.49
#> 10 10     0   NA 1.40
```

Treatment group mean = 6.75

Control group mean = 2.82

Difference-in-means = 3.93

Difference-in-means = **average treatment effect** + **selection bias**

= 2.82 + (3.93 − 2.82) = 2.82 + 1.11

Selection bias $\neq 0 \implies$ people who "select into" treatment are different

Randomized control trials

Randomized control trials

Overcoming selection bias

The problem? The existence of selection bias precludes making *other-things-equal* comparisons.

- To make valid comparisons that identify causal effects, we need to shut down the bias coming from selection.

The solution? Conduct an experiment!

- How? Assign treatment **randomly**
- Hence the name, ***randomized* control trial** (RCT)

Randomized control trials

Example: Effect of de-worming on attendance

Motivation: Intestinal worms are common among children in less-developed countries. The symptoms of these parasites can keep school-aged children at home, disrupting human capital accumulation.

Policy question: Do school-based de-worming interventions provide a cost-effective way to increase school attendance?

Randomized control trials

Example: Effect of de-worming on attendance

Research question: How much do de-worming interventions increase school attendance?

Q: Could we simply compare average attendance among children with and without access to de-worming medication?

- **A:** If we're after the causal effect, probably not. (Why not?)

Selection bias: Families with access to de-worming medication probably have healthier children for other reasons, too (wealth, access to clean drinking water, *etc.*).

- **We can't make an *all-else-equal* comparison** $\longrightarrow$ in expectation, observed differences will deviate *systematically* from the ATE!

Randomized control trials

Example: Effect of de-worming on attendance

Solution: Run an experiment.

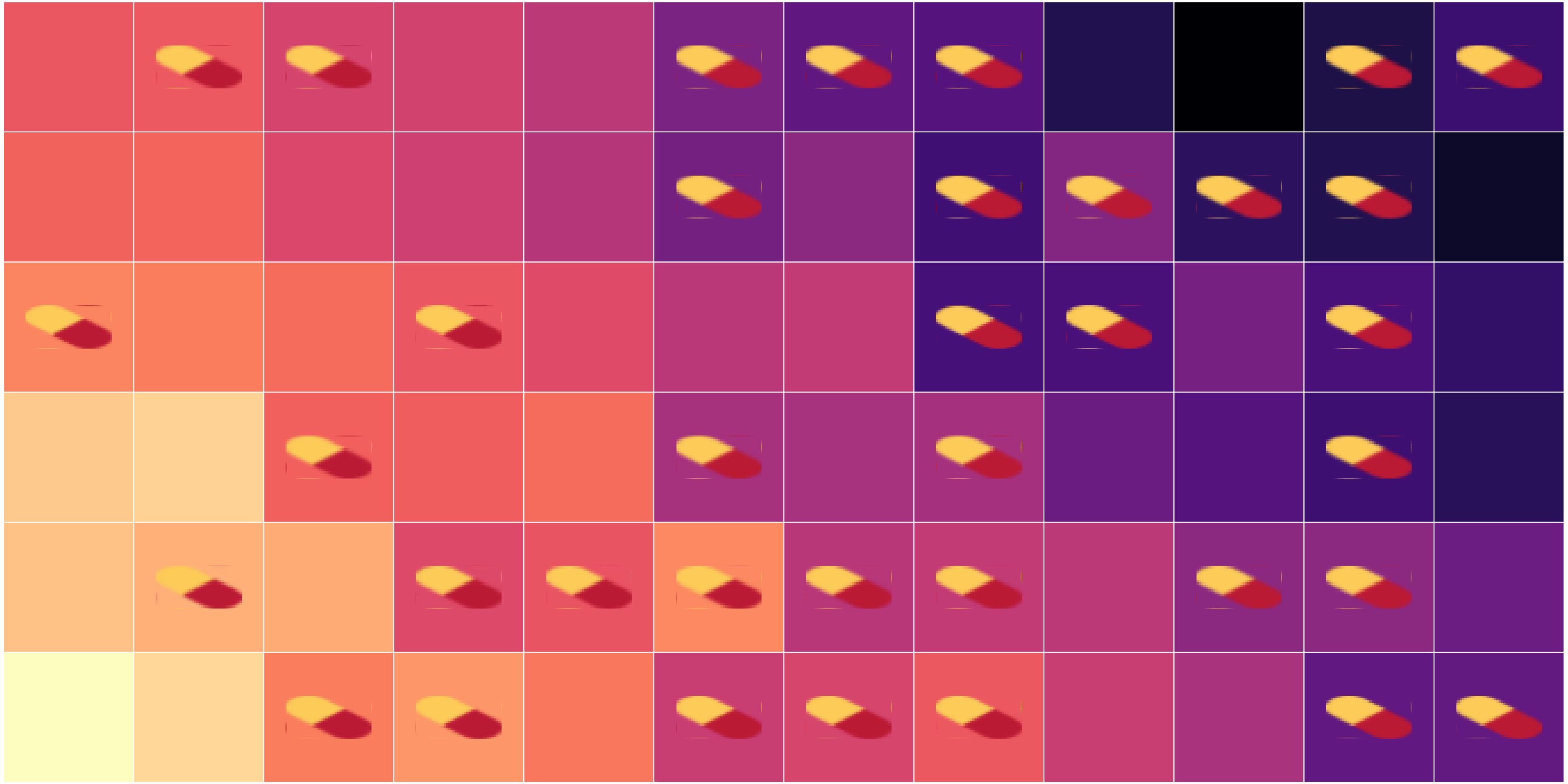
Imagine an RCT where we have two groups:

- **Treatment:** Villages where children get de-worming medication in school.
- **Control:** Villages where children don't get de-worming medication in school (status quo).

By randomizing villages into **treatment** or **control**, we will, on average, include all kinds of villages (poor *vs.* less poor, access to clean water *vs.* contaminated water, hospital *vs.* no hospital, *etc.*) in both groups.

All else equal!

72 villages of varying levels of development + randomly assigned treatment



Randomized control trials

Example: Effect of de-worming on attendance

We can estimate the **causal effect** of de-worming on school attendance by **comparing the average attendance rates** in the **treatment group** () with those in the **control group** (no):

Treatment group attendance rate – **Control group attendance rate**

Result: This was done in Kenya, where attendance increased with the random assignment of treatment.

- 25-percent decrease in absenteeism at a cost of $0.60 per child
- Long term cost effectiveness: Additional 11.91 years of schooling per $100 spent on de-worming

Source: Deworming to increase school attendance, *Abdul Latif Jameel Poverty Action Lab.*

Randomized control trials

Example: Effect of de-worming on attendance

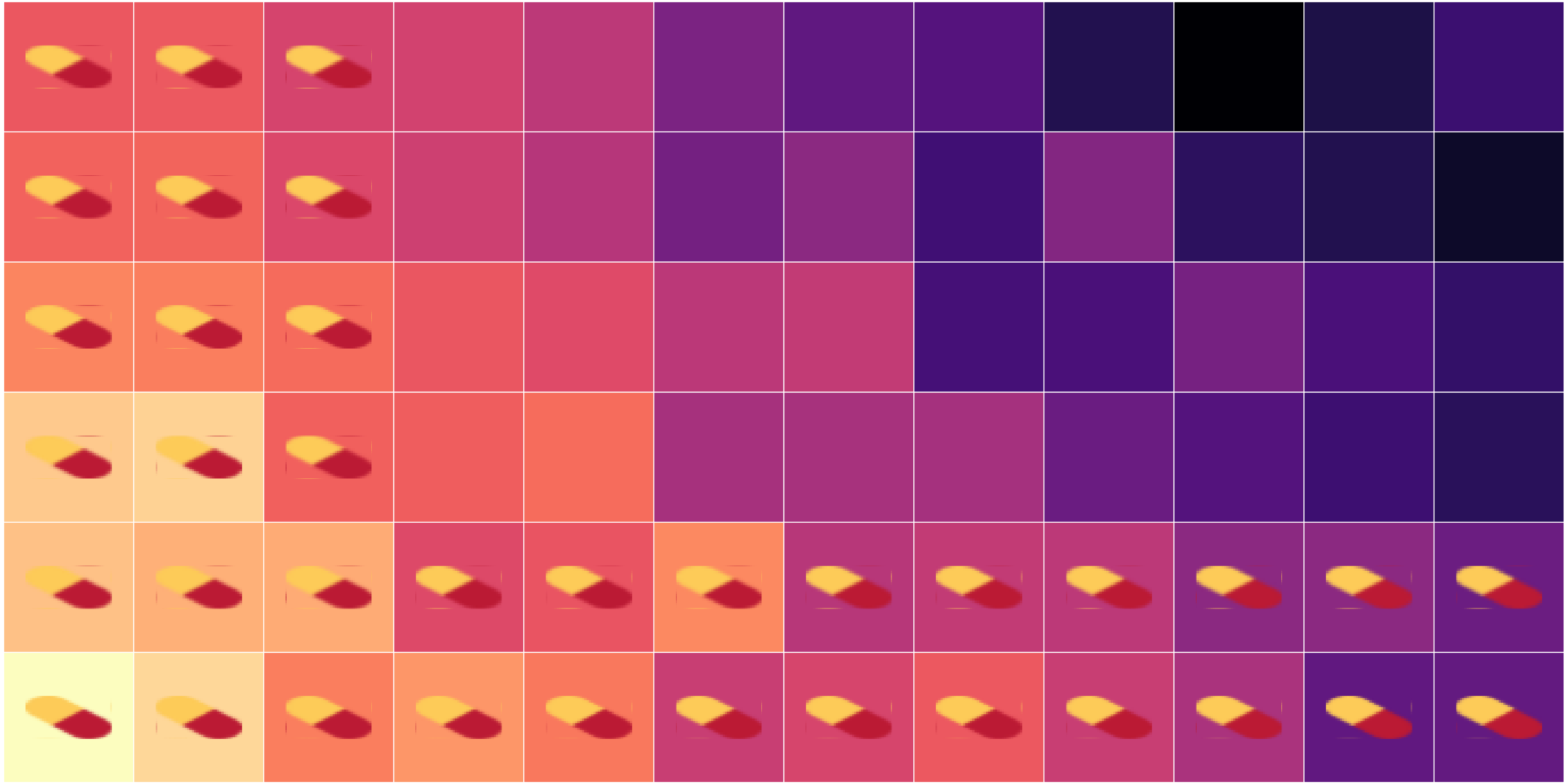
We can estimate the **causal effect** of de-worming on school attendance by **comparing the average attendance rates** in the **treatment group** () with those in the **control group** (no):

Treatment group attendance rate – **Control group attendance rate**

Q: Should we trust the results of the comparison?

A: Even with healthy skepticism, we probably should? On average, randomly assigning treatment balances the treatment and control groups across other dimensions that could explain school attendance.

Balance ***on average*** $\neq$ Balance ***every time***



Interpreting results

Internal validity

Addresses the question, ***should we believe the study?***

A study has high **internal validity** if, within the context of the study, we are confident that one variable has a **causal** influence on the outcome of interest (*e.g.*, there's **no selection bias**).

External validity

Addresses the question, ***how far can we generalize the results of the study?***

A study has high **external validity** to the extent that the results **apply to other contexts** (not just the local environment that generated the results).

Thinking Fast and Slow, Chicago edition

Thinking Fast and Slow, Chicago edition

Background

Policy question: How can we reduce violent crime among young men?

Research agenda: What factors influence an individual's proclivity toward violent crime?

- Self control? Social skills? Grit?
- Economic hardship?
- Police presesnce?
- Early chilhood education?
- Something else?

Thinking Fast and Slow, Chicago edition

Research question: Can cognitive-behavioral therapy keep young men in school and out of trouble?

- Proposed mechanism: Automaticity.

Experiment: *Becoming a Man*

4804 young men in Chicago Public Schools randomly assigned to one of two groups:

- **Treatment group:** Group cognitive-behavioral therapy program during school (once per week for 1-2 school years)
- **Control group:** No intervention

A similar experiment was also conducted in the Cook County Juvenile Temporary Detention Center.

Source: Sara B Heller et al. (2017), Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago, *The Quarterly Journal of Economics*.

Becoming a Man: Experimental results

Outcome	Control mean	Treatment mean	Effect of treatment assignment	Effect of participation
School engagement index	0	0.04	0.04	0.088
			(0.016)	(0.034)
Total arrests per youth per year	0.603	0.53	-0.073	-0.161
			(0.031)	(0.068)
Violent	0.136	0.109	-0.027	-0.06
			(0.011)	(0.024)
Property	0.069	0.072	0.003	0.006
			(0.008)	(0.018)
Drug	0.132	0.127	-0.005	-0.011
			(0.012)	(0.027)
Other	0.266	0.222	-0.044	-0.097
			(0.019)	(0.040)

Notes: 4804 observations. Standard errors in parentheses.