Applied Economics Issues

Session 4: The Credibility Revolution and Research Designs

Ewen Gallic

Master 2 DADEE – 2026

amU Aix Marseille Université

amse
école d'économie d'aix-marseille
aix-marseille school of economics

Today's session

1. **Reading debrief** (Angrist and Pischke 2010, JEP)
2. **Selection bias and research designs**: from a coin to a cutoff
3. A closer look at **difference-in-differences**
4. **Practice session**: difference-in-differences in R
5. **Wrap-up** on causal links
6. Reading assignment for **next session**

1. Reading debrief: the credibility revolution

1.1. Mapping the argument

Your turn: map the argument of Angrist and Pischke (2010)

Groups of 3–4: one part each

1. The Leamer critique and his remedies
2. Why there is less con today
3. Research design; "whither sensitivity analysis?"
4. Macro and industrial organization
5. Has the pendulum swung too far?

For your part (15 min)

- the 2–3 main **claims**, one sentence each;
- what supports each claim: evidence, example, or reasoning;
- the claim you find **least supported**.

Once you are done, let us build the map of the paper together.

The argument of the paper

The problem (1983)

- Fragile estimates, "whimsical" assumptions
- Leamer (1983, AER) remedy: sensitivity analysis
- It helped, but did not fix things.

What took the con out

- More and better data
- Fewer distractions (robust standard errors)
- And, mostly, **better research designs**.

Limits, objections

- Not yet in macro and IO
- Results are local; questions are small
- Reply: accumulate evidence (see this excellent book on that argument: Rosenbaum (2017)).

The argument of the paper

The problem (1983)

- Fragile estimates, "whimsical" assumptions
- Leamer (1983, AER) remedy: sensitivity analysis
- It helped, but did not fix things.

What took the con out

- More and better data
- Fewer distractions (robust standard errors)
- And, mostly, **better research designs**.

Limits, objections

- Not yet in macro and IO
- Results are local; questions are small
- Reply: accumulate evidence (see this excellent book on that argument: Rosenbaum (2017)).

The argument of the paper

The problem (1983)

- Fragile estimates, "whimsical" assumptions
- Leamer (1983, AER) remedy: sensitivity analysis
- It helped, but did not fix things.

What took the con out

- More and better data
- Fewer distractions (robust standard errors)
- And, mostly, **better research designs**.

Limits, objections

- Not yet in macro and IO
- Results are local; questions are small
- Reply: accumulate evidence (see this excellent book on that argument: Rosenbaum (2017)).

1.2. The con, through an example

The con, and Leamer's remedy

Leamer's diagnosis 💬 Quote

- Leamer (1983, AER): results rest on *whimsical* assumptions.
- "*hardly anyone takes anyone else's data analysis seriously.*"

How to tackle this issue?

- Leamer: **sensitivity analysis**, the range of estimates across many sets of controls.
- Angrist and Pischke (2010, JEP): "*the primary engine driving improvement has been a focus on the* ***quality of empirical research designs****.*"

Example: Does the death penalty deter murder?

A hot topic in the U.S.

- 1972: the Supreme Court halts executions (*Furman*); 1976: it allows them again (*Gregg*).
- Homicide rates were high and rising: deterrence was a key argument.

An economic question (Becker, 1968)

- A potential murderer weighs the expected punishment: probability $\times$ severity.
- But a murderer facing execution has nothing more to lose by killing witnesses.

Hence, in theory, the sign is ambiguous: only data can tell...

Ehrlich (1975): the murder supply equation

Using U.S. time series, 1933–1969, in logs to model the rate of capital murders in the population in a given year:

$$\ln\left(\frac{Q}{N}\right)_t = c + \alpha_1 \ln P_{a,t} + \alpha_2 \ln P_{c|a,t} + \alpha_3 \ln P_{e|c,t} + \beta' \ln X_t + \lambda t + v_t$$

- P_a, $P_{c|a}$, $P_{e|c}$: probabilities of arrest, conviction, **execution**, all endogenous.
- Instruments: lagged police and government spending, population, share nonwhite.
- $\hat{\alpha}_3 \approx -0.065$: each execution is associated with 7 to 8 fewer murders.

The results from this paper were cited by the Supreme Court in *Gregg* (1976).

Your turn: what would you ask Ehrlich?

You spent two sessions on IV. What would you ask Ehrlich about his instruments?

Your turn: what would you ask Ehrlich?

You spent two sessions on IV. What would you ask Ehrlich about his instruments?

- **Relevance**: why would police spending or the share nonwhite move $P_{e|c}$?
- **Exclusion**: why would they affect murders **only** through P_a, $P_{c|a}$, $P_{e|c}$?

$\Rightarrow$ Neither is argued in the paper, which corresponds to a "mechanical" use of IV, according to Angrist and Pischke (2010, JEP) (p. 8):

> "*The use of instrumental variables to uncover causal relationships was typically mechanical, with little discussion of why the instruments affected the endogenous variables of interest or why they constitute a 'good experiment.'*"

Leamer (1983): the deterrent effect is fragile

Same question, other data: 44 U.S. states in 1950.

- A simple regression of the murder rate on a large list of variables: one execution leads to 13 fewer murders, other things equal.
- Leamer suggests looking at how this estimate varies when including/excluding some regressors, corresponding to prior views on the question: a kind of **sensitivity analysis**.

	Change in murders per execution	
Prior	**min**	**max**
Right winger	−22.6	−0.9
Rational maximizer	−15.9	−10.2
Eye-for-an-eye	−28.7	+1.9
Bleeding heart	−25.6	+12.4
Crime of passion	−17.3	+4.1

"*any inference from these data about the deterrent effect of capital punishment is too fragile to be believed.*"

Donohue and Wolfers (2005): thirty years later

- U.S. vs. Canada: homicide rates move together, death penalty or not.
- Recent studies, re-estimated: small changes in sample or controls flip the sign.
- Executions are too rare to stand out from year-to-year noise.

"*We are confident that the effects are not large, but we remain unsure even of whether they are positive or negative.*"

1.3. Has there been a revolution?

Causal claims in economics: credibility revolution

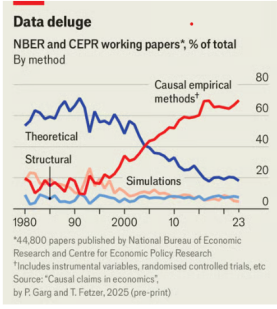


CHART: THE ECONOMIST

But even if economics is not the worst offender, it is no stranger to social science's replication crisis. Its biggest recent trend has been empirical research, with a focus on credible causal designs (see chart). Statistically significant results are prized, incentivising cherry-picking and selective presentation of results. Prashant Garg of Imperial College London and Thiemo Fetzer of the University of Warwick find that the share of papers reporting "null results" fell from 15% in 1980 to 9% in 2023. Use of private data doubled.

Causal claims, from The Economist, covering Garg and Fetzer (2025) who have analyzed over 44,000 working papers from the NBER and the CEPR.

Is there really a revolution? Counting

- **In Angrist and Pischke (2010)**: EconLit hits for "identification strategy", "research design"...:
  - **19** in 1970–1989,
  - **742** in 1990–2009.
- **Some updates**:
  - Garg and Fetzer (2025), 44,000 working papers (previous slide).
  - Goldsmith-Pinkham (2026): **In 2024**: quasi-experimental methods in 63% of applied micro papers, 39% in macro; mostly DiD.

A word of caution: as Angrist and Pischke warn, the counts track words, not good designs: "*just because the author uses the term "research design" does not mean that he or she has a good one!*"

"Whither sensitivity analysis?"

- Sala-i-Martin (1997, AER), "I just ran two million regressions": growth on 3 fixed controls, the variable tested, and every trio of the 58 others.
- 22 of 59 variables pass, among them the **fraction of the population Confucian**.
- Angrist and Pischke (2010): "*Are these the right controls? Are six controls enough?*"

$\Rightarrow$ Robustness across arbitrary controls says nothing about **where the variation comes from**.

Where the revolution had not arrived (in 2010): macro and IO

Macroeconomics

- Calibrated models: "*it's still theory.*"
- An exception: Romer and Romer (1989, NBER) read the Fed minutes to find monetary "experiments".

Industrial organization

- Mergers are **simulated** from estimated demand.
- Angrist and Pischke (2010) prefer a DiD around an actual merger (Hastings 2004, AER).

1.4. Has the pendulum swung too far?

Has the design pendulum swung too far? (1/2)

Charge 1: external validity

- Tennessee STAR: over 10,000 pupils randomly assigned to small or regular classes, kindergarten to grade 3. Small classes raise test scores (Krueger, 1999).
- But classes went from about 25 to 15 pupils, in Tennessee. What about 40 to 30, elsewhere?

Angrist and Pischke's reply

- "*Empirical evidence on any given causal effect is always local [. . .]*" 💬 Quote
- Solution: accumulating empirical evidence (varying situations, locations, etc.)

Has the design pendulum swung too far? (2/2)

Charge 2: are we still asking important questions?

- Citing Scheiber (2007): "*young economists have turned away from important questions like poverty, inequality, and unemployment to study behavior on television game shows*"
- Raj Chetty: "*People think about the question less than the method [. . .] so you get weird papers, like sanitation facilities in Native American reservations.*"

Angrist and Pischke's reply: narrow does not mean trivial

- Gym contracts reveal present-oriented bias (DellaVigna and Malmendier 2006, AER).
- Big questions too: the slave trade (Nunn 2008, QJE), climate (Deschênes and Greenstone 2007, AER).
- "*Good research designs complement good questions.*"

The replies, in the same issue of the JEP (Spring 2010)

Reply	**Main point**
Leamer (2010)	Designs rest on assumptions too: sensitivity analysis is still needed
Keane (2010)	A "*false sense of certainty*" : all inference needs untestable assumptions
Sims (2010)	Natural experiments "*are not in fact experiments*" ; on macro, "*mainly nonsense*"
Nevo and Whinston (2010)	Structural models and credible designs are "*complements, not substitutes*"
Stock (2010)	The other change: inference robust to side assumptions (robust SEs, weak instruments)

2. Selection bias: the problem

2.1. Who gets the treatment?

Does fertilizer raise yields? A farmer or a coin

Does fertilizer raise yields? Consider a field, split into plots. Some plots will receive fertilizer. Either:

1. The farmer chooses

- They fertilize their best plots.
- Fertilized plots yield more: because of the fertilizer, or of the soil?

2. A coin chooses

- Heads: the plot is fertilized.
- Good and poor soils end up in both groups, on average.

With the coin, the difference in mean yields measures the effect of the fertilizer: a **randomized controlled trial (RCT)**.

Recap (S3): two potential outcomes, one observed

Four plots. Each has two potential yields: $Y_i(1)$ with fertilizer, $Y_i(0)$ without.

	Plots	$Y_i(0)$	$Y_i(1)$	$\tau_i = Y_i(1) - Y_i(0)$
Good soil	2	10	12	2
Poor soil	2	6	8	2

- A plot is fertilized or not ($D_i = 1$ or 0): we see $Y_i = Y_i(D_i)$, **never both**.
- (Here, we can write both columns only because the example is made up.)

What does the farmer's comparison measure?

Consider the case where the farmer fertilizes the two good plots.

The difference in mean yields splits in two (add and subtract $\mathrm{E}[Y(0)|D{=}1]$):

$$\underbrace{\mathrm{E}[Y|D{=}1] - \mathrm{E}[Y|D{=}0]}_{\text{observed difference}} = \underbrace{\mathrm{E}[Y(1) - Y(0)|D{=}1]}_{\text{effect on the treated}} + \underbrace{\mathrm{E}[Y(0)|D{=}1] - \mathrm{E}[Y(0)|D{=}0]}_{\text{selection bias}}$$

With the numbers from the table in the previous slide: $12 - 6 = 2 + (10 - 6)$.

Hence, the **bias** is twice the **effect**.

In this example, this reflects the fact that good plots are **more likely to be fertilized**, and they yield more **even without fertilizer**.

Selection bias: who gets the treatment?

Selection bias: how different the two groups would have been **without** the treatment.

Treatment	**Who chooses**	**Why $Y(0)$ differs**	**Bias**
College	students, families	abler, richer	$+$
Training program	the unemployed	recent job loss	$-$
Hospital	the sick	already ill	$-$
Small class	head teachers	weaker pupils	$-$

The bias can hide an effect, or reverse its sign. For example, hospitals look harmful if we ignore this bias.

Selection bias is an omitted variable (S2)

Suppose the effect is the same for everyone (τ), and write $Y_i(0) = \alpha + u_i$. Then:

$$Y_i = \alpha + \tau D_i + u_i$$

The difference in means is the OLS slope of Y on D:

$$\mathrm{E}[Y|D{=}1] - \mathrm{E}[Y|D{=}0] \;=\; \tau \;+\; \underbrace{\mathrm{E}[u|D{=}1] - \mathrm{E}[u|D{=}0]}_{\text{selection bias}}$$

- u gathers what also determines $Y(0)$: ability, health, soil, motivation...
- Selection bias is D **correlated with the error term** (*i.e.*, the omitted variable bias of S2).

Why not simply control for it?

It depends on whether the selection is on:

- **Observables**: given the controls X, treatment is as good as random. Regression can then work.
- **Unobservables**: motivation, health, the farmer's knowledge of his soil. We cannot control for what we do not see.

2.2. A ladder of designs

Four designs, from the strongest to the weakest argument

All aim at the same thing: an **argument that the selection bias is zero**. Where does the variation come from?

Design	**Source of variation**	**Method**
1. Randomized trial	a coin	difference in means
2. Natural experiment	chance (a lottery)	difference in means; IV
3. Regression discontinuity	a rule (a cutoff)	comparison at the cutoff
4. Difference-in-differences	timing $\times$ a comparison group	change vs. change

The weaker the randomness, the more we have to **argue**.

3. Randomized trials

3.1. Field experiments

Random assignment removes the selection bias

Now, let us turn to the case where assignment is **randomized**. The coin knows nothing about the plots:

$$(Y(0), Y(1)) \perp\!\!\!\perp D \;\Rightarrow\; \mathrm{E}[Y(0)|D{=}1] = \mathrm{E}[Y(0)|D{=}0] = \mathrm{E}[Y(0)]$$

Hence:

$$\underbrace{\mathrm{E}[Y|D{=}1] - \mathrm{E}[Y|D{=}0]}_{\text{observed difference}} = \underbrace{\mathrm{E}[Y(1) - Y(0)|D{=}1]}_{\text{effect on the treated}} + \underbrace{\mathrm{E}[Y(0)|D{=}1] - \mathrm{E}[Y(0)|D{=}0]}_{\text{selection bias} = 0}$$

- **Same soils in both groups**, on average: the difference in means measures the average effect $\mathrm{E}[Y(1) - Y(0)]$ (here, 2).
- **On average, not in every sample** (Leamer, 1983): a draw can be unlucky.

$\Rightarrow$ A research design is an **argument that the selection bias is zero**, or that it can be removed.

Famous trials (1/2): cash transfers and class size

Some examples of famous RCTs:

Progresa, Mexico (Gertler 2004, AER)

- Cash to poor families if children attend school and clinics.
- 506 villages: 320 start the program, drawn at random; the others wait two years.
- Treated children: about 23% fewer illnesses, 18% less anemia.

STAR, Tennessee (Krueger 1999, QJE)

- Do smaller classes raise test scores?
- About 11,600 pupils assigned at random to small or regular classes.
- Small class: +4 percentile points in the first year.

Read them if you are interested!

Famous trials (2/2): deworming and microcredit

Two other papers to read!

Deworming, Kenya (Miguel and Kremer 2004, ECTA)

- 75 schools, treated in turn from 1998.
- Absenteeism falls by about a quarter, also in untreated schools nearby: **spillovers**.

Microcredit, India (Banerjee et al. 2015, AEJ: Applied)

- 104 slums: 52 drawn at random get a lender's branch.
- More borrowing, but no change in average consumption: no "miracle".

3.2. Experiments in the laboratory

Lab experiments: what are they for?

Lab Experiments



LABoratoire d'EXpérimentation en Economie et Management LABEX-EM, in Rennes (France). Source: CREM.

- The aim is to **test a theory**, or to measure a behavior: cooperation, risk, patience, fairness.
- The researcher builds the environment: **rules, payoffs, information**.
- Only one feature changes between treatments: any difference in behavior comes from it.
- Participants are paid according to their choices, so that incentives are real (Smith 1976, AER).

An example: cooperation and punishment

- Participants are asked to play a **game**: in a group of four, each player puts money in a common pot. Everyone gains from contributions, but each prefers that the others pay.
- The researchers want to test a **theory**: selfish players contribute nothing. Punishing others costs money, so no one punishes.
- A **treatment** is considered: with or without a stage where players can pay to punish others.
- This famous game highlighted the following **result**: free riders are heavily punished, even at a cost, and cooperation rises sharply (Fehr and Gächter 2000, AER).

Lab experiments and randomized trials

In RCTs and lab experiments, the researcher assigns the treatment. Both usually answer different questions.

	Randomized trial (field)	**Lab experiment**
Setting	real life: schools, villages	a laboratory
Participants	the people a policy targets	mostly students
Question	what does this policy do?	does the theory hold?
Strength	realism	control

Lab participants are often "WEIRD" (Henrich et al., 2010); see L'Haridon and Vieider (2019, Quant. Econ.).

4. Natural and quasi-natural experiments

4.1. Chance, or an argument

Experiments and observational studies

Experiments

The **researcher** assigns the treatment.

- RCT: by randomization.
- Lab: in a controlled environment.

Observational studies

Others decide: individuals, institutions, nature.

- **With a design**: natural and quasi-natural experiments argue that the assignment is exogenous.
- **Without**: a regression with controls assumes selection on observables only.

Angrist and Pischke (2010, JEP): "*a well-done observational study can be more credible [. . .] than a poorly executed randomized trial.*"

When randomization is impossible: natural experiments

"*It's difficult to imagine a randomized trial to evaluate the effect of immigrants on the economy of the host country*" (Angrist and Pischke 2010, JEP).

Sometimes **nature or an institution** assigns a treatment **by chance**:

- centralized school assignment in New York City and Denver: seat offers are partly random (Angrist et al. 2024, AER: Insights);
- the weather (Deschênes and Greenstone, 2011).

Many are then used as **instruments**. Their credibility rests on the "as if random" claim.

Natural vs. quasi-natural experiments

Often, nothing is *random*. What matters is whether the assignment is **exogenous**: unrelated to the potential outcomes, $(Y(0), Y(1)) \perp\!\!\!\perp D$: who gets treated has nothing to do with what their outcomes would have been, with or without the treatment. What guarantees it?

	RCT	**Natural**	**Quasi-natural**
Who assigns?	the researcher	nature, an institution	an institution, a rule
Exogeneity comes from	the coin	chance	**an argument**

Quasi-natural: not random, only **plausibly** exogenous. E.g., the 1920s immigration quotas (Abramitzky et al. 2023, AEJ: Applied), the Texas Top Ten Percent rule (Black et al. 2023, AEJ: Applied).

When it fails: who signs up for training?

- $D_i = 1$: worker i enrolls in a training program. $Y_i(0)$: their earnings next year **without** training.
- Consider cases where workers enroll after a bad year (job loss, fewer hours). Part of that loss is temporary: their earnings would recover anyway.
- In such cases, knowing that $D_i = 1$ tells us that $Y_i(0)$ is low now and about to rebound: $Y(0) \not\perp\!\!\!\perp D$.

Both simple comparisons are then biased, in opposite directions:

- **Trainees vs. non-trainees**: trainees start lower, so training looks **too weak**.
- **Trainees, before vs. after**: the rebound adds to the effect, so training looks **too strong**.

We come back to this case in the difference-in-differences part, with the **Ashenfelter dip**.

Natural experiments: a note on credibility

(Rosenbaum, 2017, p. 107)

The term 'reasonably compelling' is intended to suggest, as one must, that even a reasonably compelling observational study may turn out, in light of subsequent research, to have reached an erroneous conclusion. Sometimes a reasonably compelling observational study prompts investigators to perform a randomized trial, and sometimes the trial does not support the conclusions of the observational study. At other times, several reasonably compelling observational studies point in incompatible directions. When ethical or practical constraints force scientists to rely on observational studies, it is not uncommon to see a decade or more of thrashing about, a decade or more of controversy, conflicting conclusions, and uncertainty. This can be true even when the studies themselves are well designed and executed.

What makes an assignment "plausibly exogenous"?

- One has to ask how the rule works. Who could influence it? Why would treated and comparison units be alike, **apart from the treatment**? Answering these questions requires an **argument**.
- It is possible to run some **checks**: compare the groups **before** the treatment (balance, pre-trends), try placebo outcomes or dates, look for manipulation of a cutoff (we will see them with RD and DiD).
- However, the assumption is about potential outcomes ($Y(0)$ of the treated, $Y(1)$ of the untreated), which we never observe. Data can contradict it, but never prove it.

The next designs we are about to develop weaken the requirement in different ways:

- **Regression discontinuity**: as good as random near a cutoff.
- **Difference-in-differences**: same trend without the treatment.

5. Regression discontinuity

5.1. The idea

Rules with a cutoff

Many rules give a treatment according to a **score** X and a **threshold** τ:

- a scholarship for an exam score of 70 or more;
- Maimonides' rule: a grade of more than 40 pupils is split into two classes;
- an age threshold: a legal drinking age, a retirement age;
- an election: a candidate wins with more than 50% of the votes.

The score is the **running variable**; the threshold is the **cutoff**.

Just below, just above

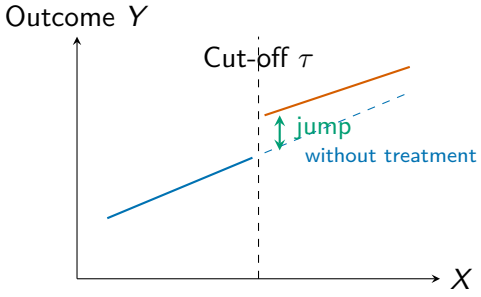
Take the scholarship (exam score X, cutoff 70, and an outcome measured after the scholarship, e.g., enrollment, graduation, or earnings).

- A student with 40 and a student with 95 differ in many ways. Comparing them mixes the effect of the scholarship with selection bias.
- A student with 69 and a student with 71 are almost identical. Yet only one gets the scholarship.
- Near the cutoff, who ends up above is **almost a matter of luck**: as good as random.

The idea is to **compare units just on each side of the cutoff**.

Regression discontinuity: the idea

- **Sharp rule**: $D = 1$ if $X \geq \tau$, $D = 0$ otherwise.
- Plot the outcome Y against X. Without the treatment, Y would follow the **same smooth line**.
- The **jump** at τ is the **effect of the treatment**, for units at the cutoff.



What we need: nothing else jumps at the cutoff

The jump in outcomes at τ identifies the effect at the cutoff:

$$\text{effect at cutoff} = \lim_{x \downarrow \tau} \mathrm{E}[Y|X{=}x] - \lim_{x \uparrow \tau} \mathrm{E}[Y|X{=}x] \;=\; \mathrm{E}[Y(1) - Y(0) \,|\, X{=}\tau]$$

if $\mathrm{E}[Y(0)|X{=}x]$ and $\mathrm{E}[Y(1)|X{=}x]$ are **continuous** at τ.

- This effect is **local**: it is estimated for units at the cutoff, not for everyone.
- The design fails if units can manipulate X to land above the cutoff.

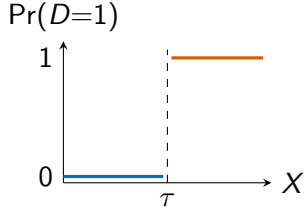
Sharp or fuzzy?

The cutoff changes the **probability of being treated**, $\Pr(D{=}1 \mid X)$

- E.g.: a score of 70 makes students *eligible* for a scholarship; D: they actually *receive* the scholarship (some decline, some win on appeal).

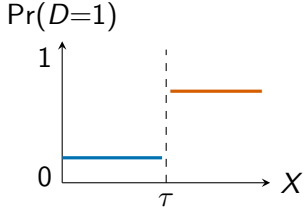
The regression discontinuity design is:

Sharp if it jumps from 0 to 1



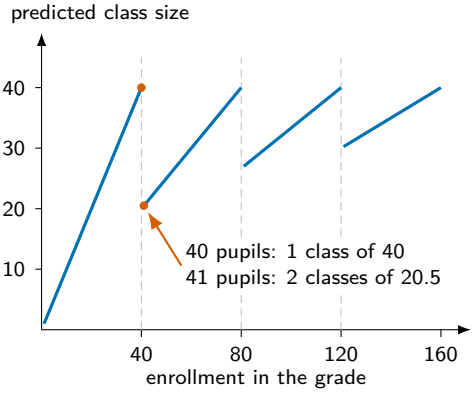
Above τ, **everyone** is treated.

Fuzzy if it jumps by less



Above τ, **more** units are treated.

Maimonides' rule: a fuzzy RD



- Angrist and Lavy (1999, QJE): in Israel, at most 40 pupils per class.
- The rule **predicts** a drop in class size at 40, 80, 120 pupils.
- Actual class size does not follow the rule exactly: this is a **fuzzy** RD design.

A fuzzy RD is an IV

Let D be class size and Y test scores. Compare schools just above and just below 40 pupils:

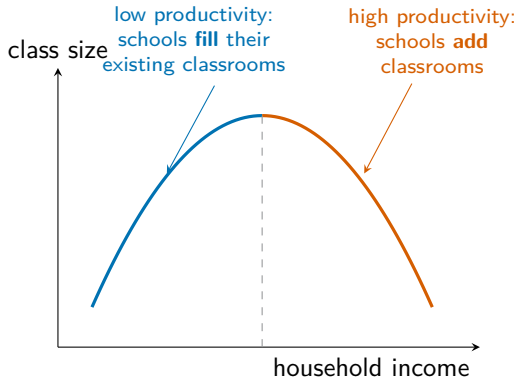
$$\text{effect of class size} = \frac{\overbrace{\lim_{x\downarrow\tau} \mathrm{E}[Y|X{=}x] - \lim_{x\uparrow\tau} \mathrm{E}[Y|X{=}x]}^{\text{jump in test scores}}}{\underbrace{\lim_{x\downarrow\tau} \mathrm{E}[D|X{=}x] - \lim_{x\uparrow\tau} \mathrm{E}[D|X{=}x]}_{\text{jump in class size}}}$$

- Being above the cutoff is the **instrument**: the Wald ratio of the reduced form to the first stage.
- If the ratio is negative: scores rise when class size falls.
- The estimation is a LATE: the effect for schools whose class size the rule moves, at the cutoff.

Regression discontinuity: what can go wrong

Regression discontinuity rests on the **key assumption** that there is no precise sorting around the cutoff.

- Urquiola and Verhoogen (2009, AER): the same kind of cap in class sizes, in **Chile**.
- Class size drops at the cutoffs, but family income **jumps**: richer families pick schools just past the cutoff.



Checking for manipulation: the density test

McCrary (2008, JoE) proposes a test of whether the density of X jumps at τ: units that sort around the cutoff pile up on the favorable side.

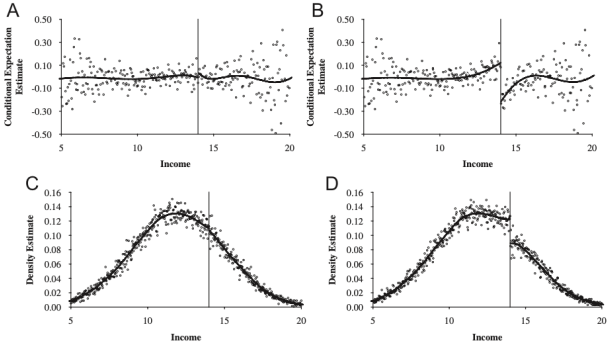


Figure 1: Hypothetical income-tested job training. **A, C**: no manipulation. **B, D**: manipulation: the density of income jumps at the cutoff, and so do the returns to treatment (Figure 2 from the paper).

Regression discontinuity: going further

The design is covered in more depth in Morgan Raux's course.

Here, we keep three things:

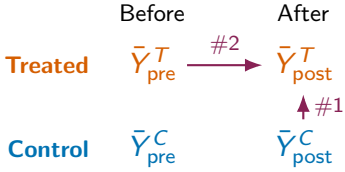
- the broad idea: compare units **just** on each side of a cutoff;
- the main assumption: **nothing else jumps** at the cutoff;
- what can go wrong: **manipulation** of the running variable.

6. Difference-in-differences

Difference-in-differences: time and groups

Let us turn to the fourth design of our ladder. The variation now comes from **timing and a comparison group**: a policy starts at a given date for some units, and not for others.

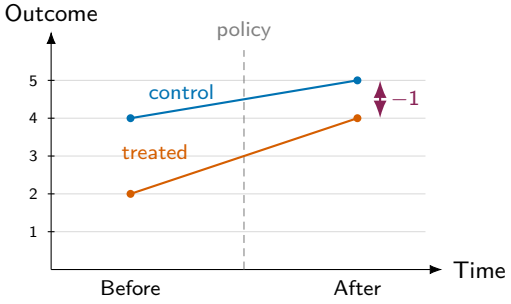
- $D_i = 1$ if unit i belongs to the **treated** group. $\bar{Y}_t^T = \mathrm{E}[Y_{it} \mid D_i{=}1]$, $\bar{Y}_t^C = \mathrm{E}[Y_{it} \mid D_i{=}0]$: mean outcome of the **treated** and **control** groups in period $t \in \{\text{pre}, \text{post}\}$.
- Target: the effect on the treated, after the policy: $\text{ATT} = \mathrm{E}[Y_{\text{post}}(1) - Y_{\text{post}}(0) \mid D{=}1]$.



Two simple comparisons are tempting, and both fail:

- #1 treated vs. control, after;
- #2 treated, before vs. after.

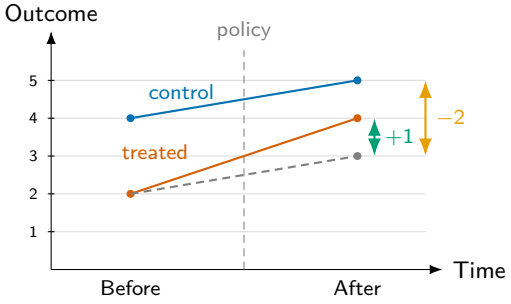
Comparison #1: treated vs. control, after the policy



- After the policy, the **treated** do worse than the **control**: $4 - 5 = -1$.
- But they were already behind **before**: the gap mixes the **effect** ($+1$) and the **selection bias** (-2), as with the fertilizer.
- Here, the policy even looks harmful.

$$\underbrace{\bar{Y}^{T}_{\text{post}} - \bar{Y}^{C}_{\text{post}}}_{-1} = \underbrace{\text{ATT}}_{+1} + \underbrace{\text{E}[Y_{\text{post}}(0) \mid D{=}1] - \text{E}[Y_{\text{post}}(0) \mid D{=}0]}_{\text{selection bias} = 3-5 = -2}$$

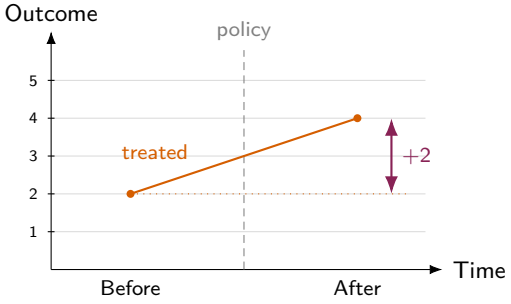
Comparison #1: treated vs. control, after the policy



- After the policy, the **treated** do worse than the **control**: $4 - 5 = -1$.
- But they were already behind **before**: the gap mixes the **effect** ($+1$) and the **selection bias** (-2), as with the fertilizer.
- Here, the policy even looks harmful.

$$\underbrace{\bar{Y}^{T}_{\text{post}} - \bar{Y}^{C}_{\text{post}}}_{-1} = \underbrace{\text{ATT}}_{+1} + \underbrace{\text{E}[Y_{\text{post}}(0) \mid D{=}1] - \text{E}[Y_{\text{post}}(0) \mid D{=}0]}_{\text{selection bias} = 3-5 = -2}$$

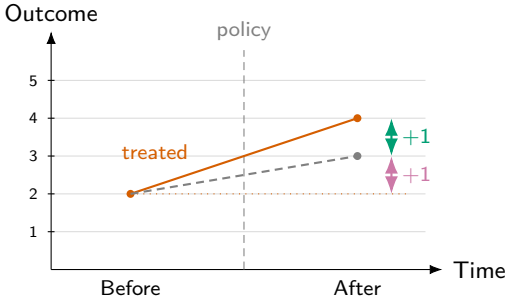
Comparison #2: the treated, before vs. after



- The **treated** do better after the policy: $4 - 2 = +2$.
- But outcomes change over time anyway (business cycle, prices, seasons): the change mixes the **effect** $(+1)$ and the **change without the policy** $(+1)$.
- Here, the effect looks twice as large.

$$\underbrace{\bar{Y}^T_{\text{post}} - \bar{Y}^T_{\text{pre}}}_{+2} = \underbrace{\text{ATT}}_{+1} + \underbrace{\text{E}[Y_{\text{post}}(0) - Y_{\text{pre}}(0) \mid D{=}1]}_{\text{change without the policy} = 3-2 = +1}$$

Comparison #2: the treated, before vs. after

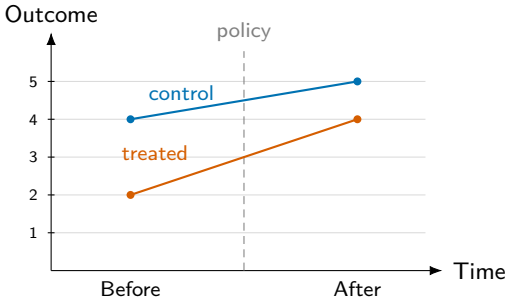


- The **treated** do better after the policy: $4 - 2 = +2$.
- But outcomes change over time anyway (business cycle, prices, seasons): the change mixes the **effect** $(+1)$ and the **change without the policy** $(+1)$.
- Here, the effect looks twice as large.

$$\underbrace{\bar{Y}^{T}_{\text{post}} - \bar{Y}^{T}_{\text{pre}}}_{+2} = \underbrace{\text{ATT}}_{+1} + \underbrace{\text{E}[Y_{\text{post}}(0) - Y_{\text{pre}}(0) \mid D{=}1]}_{\text{change without the policy} = 3-2 = +1}$$

6.1. The idea of difference-in-differences

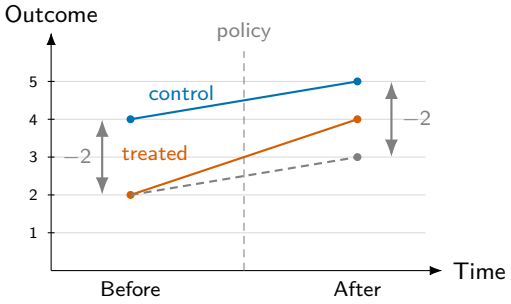
The idea: compare the changes



- **Treated**: $4 - 2 = +2$, effect and trend.
- **Control**: $5 - 4 = +1$, the trend only.
- **Parallel trends**: without the policy, the treated would have followed the trend of the control: $2 + 1 = 3$, keeping the same gap ($5 - 2 = 3$).
- The **effect**: $4 - 3 = +1$.

$$\text{DiD} = \underbrace{(\bar{Y}^T_{\text{post}} - \bar{Y}^T_{\text{pre}})}_{4-2} - \underbrace{(\bar{Y}^C_{\text{post}} - \bar{Y}^C_{\text{pre}})}_{5-4} = 1$$

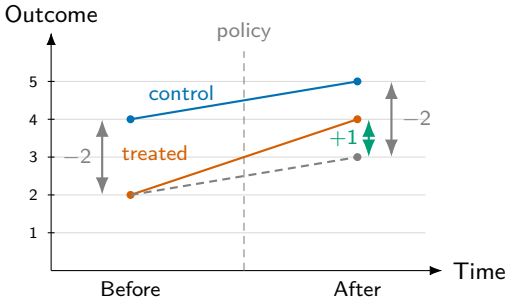
The idea: compare the changes



- **Treated**: $4 - 2 = +2$, effect and trend.
- **Control**: $5 - 4 = +1$, the trend only.
- **Parallel trends**: without the policy, the treated would have followed the trend of the control: $2 + 1 = 3$, keeping the same gap ($5 - 2 = 3$).
- The **effect**: $4 - 3 = +1$.

$$\text{DiD} = \underbrace{(\bar{Y}^T_{\text{post}} - \bar{Y}^T_{\text{pre}})}_{4-2} - \underbrace{(\bar{Y}^C_{\text{post}} - \bar{Y}^C_{\text{pre}})}_{5-4} = 1$$

The idea: compare the changes



- **Treated**: $4 - 2 = +2$, effect and trend.
- **Control**: $5 - 4 = +1$, the trend only.
- **Parallel trends**: without the policy, the treated would have followed the trend of the control: $2 + 1 = 3$, keeping the same gap ($5 - 2 = 3$).
- The **effect**: $4 - 3 = +1$.

$$\text{DiD} = \underbrace{(\bar{Y}^T_{\text{post}} - \bar{Y}^T_{\text{pre}})}_{4-2} - \underbrace{(\bar{Y}^C_{\text{post}} - \bar{Y}^C_{\text{pre}})}_{5-4} = 1$$

Difference-in-differences: the equations

Let $\Delta Y_i(0) = Y_{i,\text{post}}(0) - Y_{i,\text{pre}}(0)$: the change of unit i **without** the policy.

- **Treated** (comparison #2): $\bar{Y}^T_{\text{post}} - \bar{Y}^T_{\text{pre}} = \text{ATT} + \mathrm{E}[\Delta Y(0) \mid D{=}1]$
- **Control**, never treated: $\bar{Y}^C_{\text{post}} - \bar{Y}^C_{\text{pre}} = \mathrm{E}[\Delta Y(0) \mid D{=}0]$

Taking the difference:

$$\text{DiD} = \text{ATT} + \underbrace{\mathrm{E}[\Delta Y(0) \mid D{=}1] - \mathrm{E}[\Delta Y(0) \mid D{=}0]}_{\text{difference in trends without the policy}}$$

Parallel trends

$\mathrm{E}[\Delta Y(0) \mid D{=}1] = \mathrm{E}[\Delta Y(0) \mid D{=}0]$: *without the policy, both groups would have changed alike. Then DiD = ATT.*

Parallel trends: weaker than random assignment

Randomized trial

Same **levels** without the treatment:

$$\mathrm{E}[Y(0) \mid D{=}1] = \mathrm{E}[Y(0) \mid D{=}0]$$

Difference-in-differences

Same **changes** without the treatment:

$$\mathrm{E}[\Delta Y(0) \mid D{=}1] = \mathrm{E}[\Delta Y(0) \mid D{=}0]$$

- The groups may differ in levels: the gap is differenced out.
- It is about $Y_{\text{post}}(0)$ of the treated, never observed: it **cannot be tested**, only made plausible.
- It fails if something else hits the treated at the same time, or if units are treated **because** of their recent outcomes.
- It also assumes no effect before the policy starts (no anticipation).

Two ways to read the same number

	Before	After	After − Before
Treated	2	4	+2 (#2)
Control	4	5	+1
Treated − Control	−2	−1 (#1)	+1

- **By rows**: comparison #2 (+2), minus the trend of the control (+1).
- **By columns**: comparison #1 (−1), minus the gap that existed before (−2).

$$(\bar{Y}^T_{\text{post}} - \bar{Y}^T_{\text{pre}}) - (\bar{Y}^C_{\text{post}} - \bar{Y}^C_{\text{pre}}) = (\bar{Y}^T_{\text{post}} - \bar{Y}^C_{\text{post}}) - (\bar{Y}^T_{\text{pre}} - \bar{Y}^C_{\text{pre}}) = +1$$

6.2. An example: the minimum wage in New Jersey

Does a higher minimum wage destroy jobs?

Does a higher minimum wage reduce employment? In a competitive labor market, it should.

- Card and Krueger (1994, AER) investigate this question with a policy change: on April 1, 1992, New Jersey raises its minimum wage from \$4.25 to \$5.05. Neighboring Pennsylvania does not.
- The paper uses data on 410 fast-food restaurants, surveyed in Feb.–March 1992 (**before**) and Nov.–Dec. 1992 (**after**).

Who is treated? Who is the control group? What would comparisons #1 and #2 miss? Let us see!

The minimum wage in New Jersey: four averages

Employment per store (FTE)	**Before**	**After**	**Change**
Pennsylvania (control)	23.33	21.17	−2.16
New Jersey (treated)	20.44	21.03	+0.59
NJ − PA	−2.89	−0.14	+2.76

- #1, after only: −0.14, "no effect". But NJ stores were already smaller (−2.89).
- #2, NJ only: +0.59. But PA lost 2.16 FTE over the same months.
- DiD: $0.59 - (-2.16) = -0.14 - (-2.89) \approx 2.76$: employment **rose** in NJ, relative to PA.

Note: FTE = full-time equivalent.

The minimum wage in New Jersey: four averages

Employment per store (FTE)	Before	After	Change
Pennsylvania (control)	23.33	21.17	−2.16
New Jersey (treated)	20.44	21.03	+0.59
NJ − PA	−2.89	−0.14	+2.76

- #1, after only: −0.14, "no effect". But NJ stores were already smaller (−2.89).
- #2, NJ only: +0.59. But PA lost 2.16 FTE over the same months.
- DiD: $0.59 - (-2.16) = -0.14 - (-2.89) \approx 2.76$: employment **rose** in NJ, relative to PA.

Note: FTE = full-time equivalent.

The minimum wage in New Jersey: four averages

Employment per store (FTE)	Before	After	Change
Pennsylvania (control)	23.33	21.17	−2.16
New Jersey (treated)	20.44	21.03	+0.59
NJ − PA	−2.89	−0.14	+2.76

- #1, after only: −0.14, "no effect". But NJ stores were already smaller (−2.89).
- #2, NJ only: +0.59. But PA lost 2.16 FTE over the same months.
- DiD: $0.59 - (-2.16) = -0.14 - (-2.89) \approx 2.76$: employment **rose** in NJ, relative to PA.

Note: FTE = full-time equivalent.

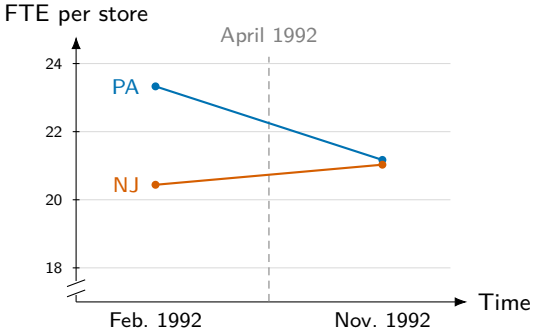
The minimum wage in New Jersey: four averages

Employment per store (FTE)	**Before**	**After**	**Change**
Pennsylvania (control)	23.33	21.17	−2.16
New Jersey (treated)	20.44	21.03	+0.59
NJ − PA	−2.89	−0.14	+2.76

- #1, after only: −0.14, "no effect". But NJ stores were already smaller (−2.89).
- #2, NJ only: +0.59. But PA lost 2.16 FTE over the same months.
- DiD: $0.59 - (-2.16) = -0.14 - (-2.89) \approx 2.76$: employment **rose** in NJ, relative to PA.

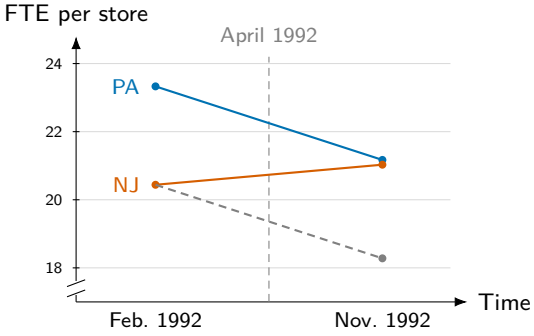
Note: FTE = full-time equivalent.

The minimum wage in New Jersey, graphically



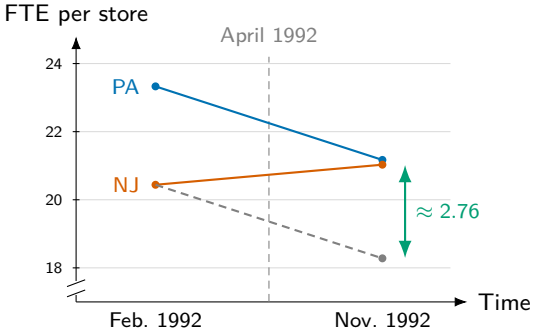
- **PA** shows what happens without the reform: -2.16 FTE per store.
- Parallel trends: NJ would have lost as much, $20.44 - 2.16 = 18.28$.
- Effect: $21.03 - 18.28 \approx 2.76$ FTE per store: **more** jobs, not fewer.

The minimum wage in New Jersey, graphically



- **PA** shows what happens without the reform: -2.16 FTE per store.
- Parallel trends: NJ would have lost as much, $20.44 - 2.16 = 18.28$.
- Effect: $21.03 - 18.28 \approx 2.76$ FTE per store: **more** jobs, not fewer.

The minimum wage in New Jersey, graphically



- **PA** shows what happens without the reform: -2.16 FTE per store.
- Parallel trends: NJ would have lost as much, $20.44 - 2.16 = 18.28$.
- Effect: $21.03 - 18.28 \approx 2.76$ FTE per store: **more** jobs, not fewer.

Exercise: what is behind each average?

These averages can be written as a regression:

$$Y_{it} = \beta_0 + \beta_1 \, \text{NJ}_i + \beta_2 \, \text{Post}_t + \beta_3 \, (\text{NJ}_i \times \text{Post}_t) + \varepsilon_{it}$$

	Before ($\text{Post}_t = 0$)	After ($\text{Post}_t = 1$)
New Jersey	$20.44 = \, ?$	$21.03 = \, ?$
Pennsylvania	$23.33 = \, ?$	$21.17 = \, ?$

Set NJ_i and Post_t to 0 or 1 in the equation: what does it say about each average?

Exercise: what is behind each average?

These averages can be written as a regression:

$$Y_{it} = \beta_0 + \beta_1 \, \text{NJ}_i + \beta_2 \, \text{Post}_t + \beta_3 \, (\text{NJ}_i \times \text{Post}_t) + \varepsilon_{it}$$

	Before ($\text{Post}_t = 0$)	After ($\text{Post}_t = 1$)
New Jersey	$20.44 = ?$	$21.03 = ?$
Pennsylvania	$23.33 = \beta_0$	$21.17 = ?$

Set NJ_i and Post_t to 0 or 1 in the equation: what does it say about each average?

Exercise: what is behind each average?

These averages can be written as a regression:

$$Y_{it} = \beta_0 + \beta_1 \, \text{NJ}_i + \beta_2 \, \text{Post}_t + \beta_3 \, (\text{NJ}_i \times \text{Post}_t) + \varepsilon_{it}$$

	Before ($\text{Post}_t = 0$)	After ($\text{Post}_t = 1$)
New Jersey	$20.44 = \beta_0 + \beta_1$	$21.03 = ?$
Pennsylvania	$23.33 = \beta_0$	$21.17 = ?$

Set NJ_i and Post_t to 0 or 1 in the equation: what does it say about each average?

Exercise: what is behind each average?

These averages can be written as a regression:

$$Y_{it} = \beta_0 + \beta_1 \, \text{NJ}_i + \beta_2 \, \text{Post}_t + \beta_3 \, (\text{NJ}_i \times \text{Post}_t) + \varepsilon_{it}$$

	Before ($\text{Post}_t = 0$)	After ($\text{Post}_t = 1$)
New Jersey	$20.44 = \beta_0 + \beta_1$	$21.03 = ?$
Pennsylvania	$23.33 = \beta_0$	$21.17 = \beta_0 + \beta_2$

Set NJ_i and Post_t to 0 or 1 in the equation: what does it say about each average?

Exercise: what is behind each average?

These averages can be written as a regression:

$$Y_{it} = \beta_0 + \beta_1 \, \text{NJ}_i + \beta_2 \, \text{Post}_t + \beta_3 \, (\text{NJ}_i \times \text{Post}_t) + \varepsilon_{it}$$

	Before ($\text{Post}_t = 0$)	After ($\text{Post}_t = 1$)
New Jersey	$20.44 = \beta_0 + \beta_1$	$21.03 = \beta_0 + \beta_1 + \beta_2 + \beta_3$
Pennsylvania	$23.33 = \beta_0$	$21.17 = \beta_0 + \beta_2$

Set NJ_i and Post_t to 0 or 1 in the equation: what does it say about each average?

Reminder: the coefficients are derivatives

$$\mathbb{E}[Y_{it} \mid \text{NJ}_i, \text{Post}_t] = \beta_0 + \beta_1 \, \text{NJ}_i + \beta_2 \, \text{Post}_t + \beta_3 \, \text{NJ}_i \, \text{Post}_t$$

For dummy variables, a "derivative" is the difference between the values at 1 and at 0.

- $\left.\frac{\partial \, \mathbb{E}[Y]}{\partial \, \text{NJ}}\right|_{\text{Post}=0} = \beta_1$: the gap between the states **before** ($20.44 - 23.33 = -2.89$).
- $\left.\frac{\partial \, \mathbb{E}[Y]}{\partial \, \text{Post}}\right|_{\text{NJ}=0} = \beta_2$: the change in Pennsylvania ($21.17 - 23.33 = -2.16$).
- $\frac{\partial^2 \, \mathbb{E}[Y]}{\partial \, \text{NJ} \, \partial \, \text{Post}} = \beta_3$: how much the change over time differs between the states, which is the **DiD** (≈ 2.76).

The regression gives the same number as the table, together with a standard error.

Two-way fixed effects: restaurant and time

With i the restaurant and t the wave (before/after the reform), the regression of Card and Krueger (1994, AER) can also be written with **two-way fixed effects** (TWFE), one per restaurant and one per wave:

$$\underbrace{Y_{it}}_{\text{employment}} = \beta \left(\underbrace{\text{Post}_t}_{\text{after}} \times \underbrace{\text{NJ}_i}_{\text{treated}}\right) + \underbrace{\gamma_i}_{\text{restaurant FE}} + \underbrace{\delta_t}_{\text{time FE}} + \varepsilon_{it}$$

- γ_i replaces $\beta_0 + \beta_1 \text{NJ}_i$: each restaurant has its own level.
- δ_t replaces $\beta_2 \text{Post}_t$: the change common to all restaurants.
- The β **coefficient** is the DiD estimate of the effect of NJ's minimum wage increase: on a balanced panel, the same number as β_3 in the 2×2 regression.

Same estimate, but which standard error? See the appendix.

Another DiD: the death penalty

Angrist and Pischke (2010, JEP) show a figure with homicide rates, U.S. vs. Canada, since 1950, from Donohue and Wolfers (2005).

- Which group is treated? What is the treatment?
- Canada is always lower: is it a problem?

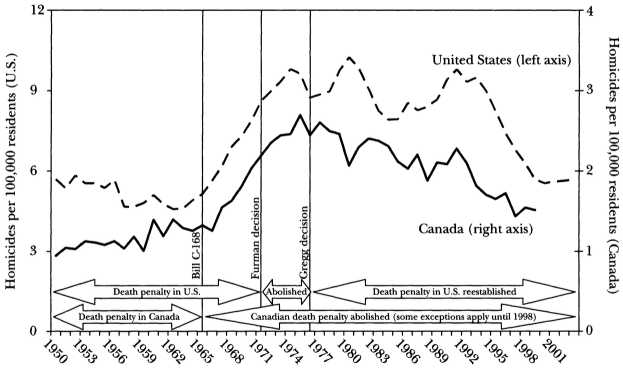


Figure 2: Homicide Rates and the Death Penalty in the United States and Canada (U.S.A. and Canada rates on the left and right y-axis, respectively). Source: Donohue and Wolfers (2005)

Another DiD: the death penalty

Angrist and Pischke (2010, JEP) show a figure with homicide rates, U.S. vs. Canada, since 1950, from Donohue and Wolfers (2005).

- Which group is treated? What is the treatment?
- Canada is always lower: is it a problem?

The U.S. suspended (1972) and restored (1976) the death penalty: the two series still move in parallel.

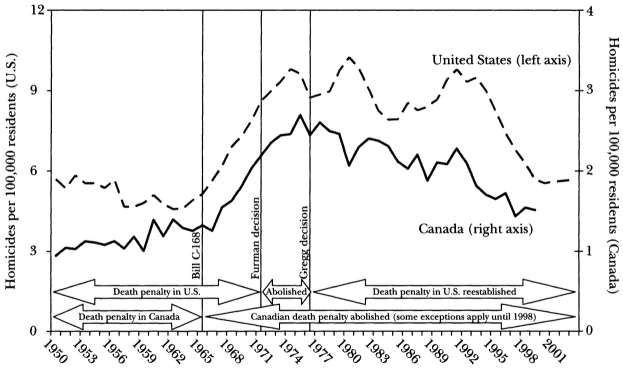


Figure 2: Homicide Rates and the Death Penalty in the United States and Canada (U.S.A. and Canada rates on the left and right y-axis, respectively). Source: Donohue and Wolfers (2005)

6.3. What can go wrong: the Ashenfelter dip

When are the trends parallel? A case study

Difference-in-differences needs **parallel trends**: without the program, treated and comparison units would have moved alike.

This is at risk when units enter the program **because of what happened to them recently**.

A case you are already familiar with: **trainees entering a training program**

- Are their earnings trends parallel to those of non-trainees?

Ashenfelter and Card (1985), seen from 2010

You read the introduction of Ashenfelter and Card (1985, REStat) a few weeks ago.

The paper you read for this session places it at the start of the credibility revolution:

> Angrist and Pischke (2010, JEP), p. 5
>
> *[...] Ashenfelter and Card (1985), using longitudinal data to evaluate federal training programs without the benefit of a quasi-experimental research design, found it difficult to construct specification-robust estimates.*

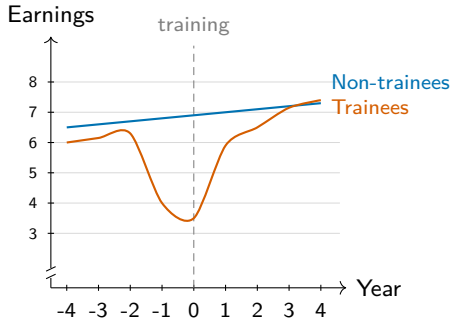
Why was it so hard to get robust estimates from earnings histories? Let us see!

A refresher: Ashenfelter and Card (1985)

Those who read Ashenfelter and Card (1985, REStat) in depth: explain it to the class before we look at the slides.

1. What happens to trainees' earnings **just before** they enter the program? Why?
2. Why does this make a simple comparison misleading?
3. How do the authors use the **longitudinal structure** of the data to deal with it?

The Ashenfelter dip: the empirical puzzle



This figure is for illustrative purposes only, the numbers are made-up.

- Empirically, earnings of workers entering training programs often **fall in the year or two** before training (job loss, layoffs).
- This creates a dip in trainees' earnings profile, called the **"Ashenfelter dip"**.

⇒ Any comparison strategy has to account for this dip; otherwise, the estimated effect of the program can be badly biased.

$\bar{Y}_t^T$ ($\bar{Y}_t^C$): mean earnings of trainees (non-trainees) in year t, training at $t = 0$. The dip: $\bar{Y}_{-1}^T < \bar{Y}_{-2}^T$, while $\bar{Y}_{-1}^C > \bar{Y}_{-2}^C$.

The Ashenfelter dip: all the numbers

Year k	-4	-3	-2	-1	**0**	1	2	3	4
Non-trainees: $\bar{Y}_k^C$	6.5	6.6	6.7	6.8	6.9	7.0	7.1	7.2	7.3
Trainees: $\bar{Y}_k^T$	6.0	6.15	6.3	4.0	3.5	5.9	6.5	7.15	7.4
Trainees, no training: $\mathrm{E}[Y_k(0) \mid D{=}1]$	6.0	6.15	6.3	4.0	3.5	5.4	6.0	6.65	6.9
Effect: ATT_k	0	0	0	0	0	0.5	0.5	0.5	0.5
Gap: $\bar{Y}_k^T - \bar{Y}_k^C$	-0.5	-0.45	-0.4	-2.8	-3.4	-1.1	-0.6	-0.05	0.1
Event study (ref. -2): β_k	-0.1	-0.05	0	-2.4	-3.0	-0.7	-0.2	0.35	0.5
Parallel path from -2: $\bar{Y}_{-2}^T + (\bar{Y}_k^C - \bar{Y}_{-2}^C)$	–	–	6.3	6.4	6.5	6.6	6.7	6.8	6.9

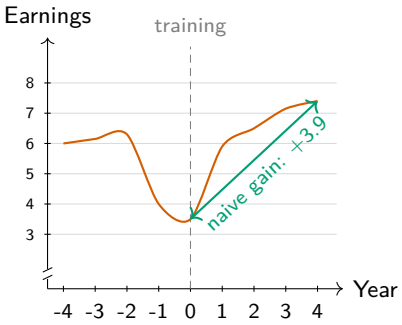
Two naive comparisons, two opposite errors

#1: after training, cross-section



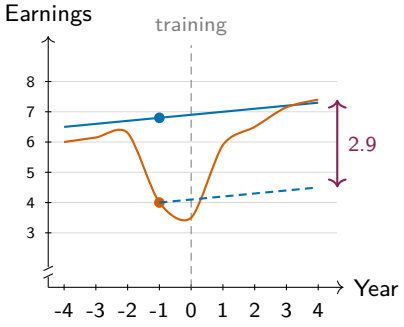
- $\bar{Y}_1^T - \bar{Y}_1^C = 5.9 - 7.0 = -1.1$: trainees still earn less, the effect looks **small**.
- But the gap is effect + selection bias (the dip).

#2: before/after, trainees only



- $\bar{Y}_4^T - \bar{Y}_0^T = 7.4 - 3.5 = +3.9$: from the bottom of the dip, the gain looks **large**.
- But the gain is effect + **recovery from the dip**.

DiD fails too, with a base year in the dip

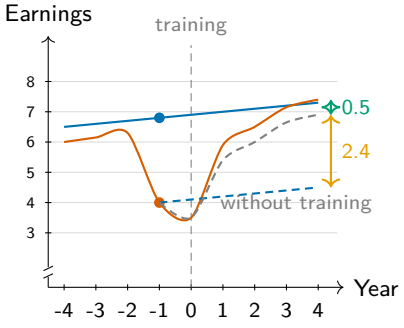


- Base year -1:
  $(\bar{Y}_4^T - \bar{Y}_{-1}^T) - (\bar{Y}_4^C - \bar{Y}_{-1}^C) =$
  $(7.4 - 4.0) - (7.3 - 6.8) = 2.9$.
- It assumes that, from the bottom of the dip, trainees would have gained $+0.5$, like non-trainees (**blue dashed**).
- Without training, they would have **recovered**: $4.0 \to 6.9$, i.e. $+2.9$ (gray).
  From year -1, the trends are **not parallel**.

$$\underbrace{2.9}_{\text{DiD}} = \underbrace{0.5}_{\text{ATT}} + \underbrace{\mathrm{E}[\Delta Y(0) \mid D{=}1] - \mathrm{E}[\Delta Y(0) \mid D{=}0]}_{(6.9-4.0)-(7.3-6.8)=2.4}$$

Here, $\Delta Y(0)$ is the change between years -1 and 4.

DiD fails too, with a base year in the dip

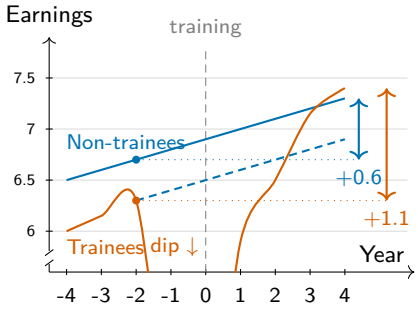


- Base year -1:
  $(\bar{Y}_4^T - \bar{Y}_{-1}^T) - (\bar{Y}_4^C - \bar{Y}_{-1}^C) =$
  $(7.4 - 4.0) - (7.3 - 6.8) = 2.9$.
- It assumes that, from the bottom of the dip, trainees would have gained $+0.5$, like non-trainees (**blue dashed**).
- Without training, they would have **recovered**: $4.0 \to 6.9$, i.e. $+2.9$ (gray).
  From year -1, the trends are **not parallel**.

$$\underbrace{2.9}_{\text{DiD}} = \underbrace{0.5}_{\text{ATT}} + \underbrace{\mathrm{E}[\Delta Y(0) \mid D{=}1] - \mathrm{E}[\Delta Y(0) \mid D{=}0]}_{(6.9-4.0)-(7.3-6.8)=2.4}$$

Here, $\Delta Y(0)$ is the change between years -1 and 4.

The fix: a base year before the dip



- Choose a base year **before** the dip: year -2.
- From -2 to 4, trainees would plausibly have grown like non-trainees: **parallel trends** is credible again.
- **But where does the dip start?** Look at every year.

$$\text{DiD} = \underbrace{(\bar{Y}_4^T - \bar{Y}_{-2}^T)}_{\Delta \text{ trainees} = 1.1} - \underbrace{(\bar{Y}_4^C - \bar{Y}_{-2}^C)}_{\Delta \text{ non-trainees} = 0.6} = 0.5$$

6.4. Looking at every year: the event study

Why not stop at one DiD number?

- The DiD rests on **parallel trends**: it concerns the counterfactual, so it cannot be tested directly. But if the groups moved in parallel **before** training, it is more credible.
- A single DiD estimate depends on the years we compare:
  - 0.5 with years -2 and 4;
  - about 1.1 with all years, in a two-way FE regression (eq. elow);
  - with all years, the "before" average includes the dip (-1, 0): the trainees look worse than usual, so the effect looks larger. Which one to believe?
- The effect does not have to be immediate or constant: one number averages effects that change over time.

$$Y_{it} = \gamma_i + \delta_t + \beta\, D_{it} + \varepsilon_{it}, \qquad D_{it} = \text{Trainee}_i \times \text{Post}_t, \qquad \text{Post}_t = \mathbb{1}_{[t \geq 1]}$$

What can be done? Estimate the DiD **year by year**, relative to one reference year, and look at the whole path.

The event study: one coefficient per year

Instead of a single DiD, estimate the DiD of **each year** with respect to a reference year:

$$Y_{it} = \underbrace{\gamma_i}_{\text{worker FE}} + \underbrace{\delta_t}_{\text{year FE}} + \sum_{k \neq -2} \beta_k \underbrace{\mathbb{1}_{[t=k]} \times \text{Trainee}_i}_{\text{trainees, in year } k} + \varepsilon_{it}$$

- β_k: how much the gap between trainees and non-trainees has changed between the **reference year** -2 and year k.
- The reference year is left out, so $\beta_{-2} = 0$ by construction. One year must be left out: otherwise, the interactions would add up to Trainee_i, already in the worker fixed effects.
- This makes us estimate **one** regression on all years (with a balanced panel, each $\hat{\beta}_k$ equals the 2×2 DiD between years -2 and k).

Each coefficient is a gap, minus the gap in year -2

Year k	-4	-3	-2	-1	0	1	2	3	4
Trainees $\bar{Y}_k^T$	6.0	6.15	6.3	4.0	3.5	5.9	6.5	7.15	7.4
Non-trainees $\bar{Y}_k^C$	6.5	6.6	6.7	6.8	6.9	7.0	7.1	7.2	7.3
Gap $\bar{Y}_k^T - \bar{Y}_k^C$	-0.5	-0.45	-0.4	-2.8	-3.4	-1.1	-0.6	-0.05	0.1
$\hat{\beta}_k = \text{gap}_k - (\textbf{-0.4})$	-0.1	-0.05	**0**	-2.4	-3.0	-0.7	-0.2	0.35	0.5

For instance:

$$\hat{\beta}_{-1} = (4.0 - 6.8) - (6.3 - 6.7) = -2.4 \qquad \hat{\beta}_4 = (7.4 - 7.3) - (6.3 - 6.7) = 0.5$$

Each $\hat{\beta}_k$ is a **2×2 DiD**: years -2 and k (i.e, 2 years), trainees and non-trainees (i.e., 2 groups).

Each coefficient is a gap, minus the gap in year -2

Year k	-4	-3	-2	-1	0	1	2	3	4
Trainees $\bar{Y}_k^T$	6.0	6.15	6.3	4.0	3.5	5.9	6.5	7.15	7.4
Non-trainees $\bar{Y}_k^C$	6.5	6.6	6.7	6.8	6.9	7.0	7.1	7.2	7.3
Gap $\bar{Y}_k^T - \bar{Y}_k^C$	-0.5	-0.45	-0.4	-2.8	-3.4	-1.1	-0.6	-0.05	0.1
$\hat{\beta}_k = \text{gap}_k - (\textbf{-0.4})$	-0.1	-0.05	**0**	-2.4	-3.0	-0.7	-0.2	0.35	0.5

For instance:

$$\hat{\beta}_{-1} = (4.0 - 6.8) - (6.3 - 6.7) = -2.4 \qquad \hat{\beta}_4 = (7.4 - 7.3) - (6.3 - 6.7) = 0.5$$

Each $\hat{\beta}_k$ is a **2×2 DiD**: years -2 and k (i.e, 2 years), trainees and non-trainees (i.e., 2 groups).

Each coefficient is a gap, minus the gap in year -2

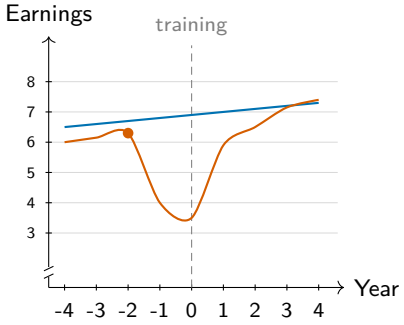
Year k	-4	-3	-2	-1	0	1	2	3	4
Trainees $\bar{Y}_k^T$	6.0	6.15	6.3	4.0	3.5	5.9	6.5	7.15	7.4
Non-trainees $\bar{Y}_k^C$	6.5	6.6	6.7	6.8	6.9	7.0	7.1	7.2	7.3
Gap $\bar{Y}_k^T - \bar{Y}_k^C$	-0.5	-0.45	-0.4	-2.8	-3.4	-1.1	-0.6	-0.05	0.1
$\hat{\beta}_k = \text{gap}_k - (\textbf{-0.4})$	-0.1	-0.05	**0**	-2.4	-3.0	-0.7	-0.2	0.35	0.5

For instance:

$$\hat{\beta}_{-1} = (4.0 - 6.8) - (6.3 - 6.7) = -2.4 \qquad \hat{\beta}_4 = (7.4 - 7.3) - (6.3 - 6.7) = 0.5$$

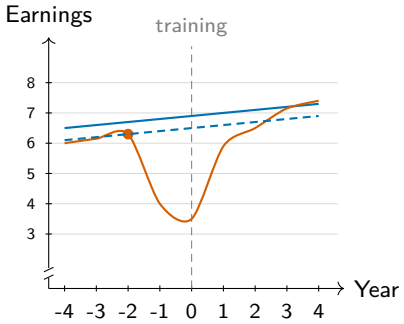
Each $\hat{\beta}_k$ is a **2×2 DiD**: years -2 and k (i.e, 2 years), trainees and non-trainees (i.e., 2 groups).

The coefficients, on the graph



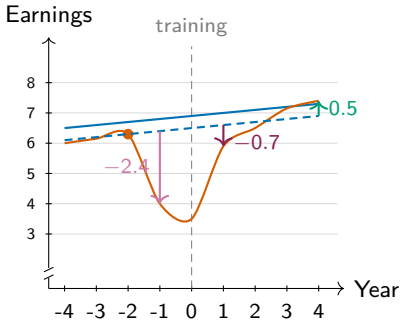
- **Blue dashed**: the non-trainees' curve, moved down by the gap of year -2 (0.4). It meets the trainees in year -2.
- $\hat{\beta}_k$ is the **vertical distance** from the dashed curve to the trainees, in year k.
- If the trends were parallel, the trainees would stay on the dashed curve: all $\hat{\beta}_k = 0$.

The coefficients, on the graph



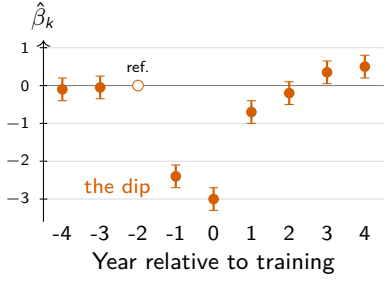
- **Blue dashed**: the non-trainees' curve, moved down by the gap of year -2 (0.4). It meets the trainees in year -2.
- $\hat{\beta}_k$ is the **vertical distance** from the dashed curve to the trainees, in year k.
- If the trends were parallel, the trainees would stay on the dashed curve: all $\hat{\beta}_k = 0$.

The coefficients, on the graph



- **Blue dashed**: the non-trainees' curve, moved down by the gap of year -2 (0.4). It meets the trainees in year -2.
- $\hat{\beta}_k$ is the **vertical distance** from the dashed curve to the trainees, in year k.
- If the trends were parallel, the trainees would stay on the dashed curve: all $\hat{\beta}_k = 0$.

Reading the event study



Illustrative, made-up numbers. Bars: 95% confidence intervals.

- **Before training** ($k < 0$): a placebo test. Parallel trends imply $\beta_k = 0$: at $k = -4, -3$, the confidence intervals **contain 0**.
- At $k = -1, 0$, trainees' earnings fall **before** the program: the trends are **not parallel** just before training. Hence the reference year -2.
- **After training** ($k > 0$): the path of the effect. At $k = 1, 2, 3$, the estimates still mix the effect and the recovery from the dip; $\hat{\beta}_4 = 0.5$, **if** the recovery is over (next slide).

The intervals are only as good as the standard errors behind them: see "Standard errors: cluster by worker".

The limit: effect, or recovery?

After training, each $\hat{\beta}_k$ mixes two things: **the effect** + **what is left of the dip**. Infinitely many stories fit the **same** coefficients; here are two:

Year k	1	2	3	4
$\hat{\beta}_k$ (the data)	-0.7	-0.2	0.35	0.5
A: effect 0.5, fast recovery	$0.5 - 1.2$	$0.5 - 0.7$	$0.5 - 0.15$	$0.5 + 0$
B: effect 1.0, slow recovery	$1.0 - 1.7$	$1.0 - 1.2$	$1.0 - 0.65$	$1.0 - 0.5$

- The data cannot tell A from B. **Even** $\hat{\beta}_4$ measures the effect only if the recovery is over by year 4. You have to **argue** it (how long do such dips last?): the data cannot show it.
- What can help: a comparison group that also had a bad year but was not trained; a longer horizon after training.

And the choice of the reference year?

Year k	-4	-3	-2	-1	0	1	2	3	4
Ref. -4	**0**	0.05	0.1	-2.3	-2.9	-0.6	-0.1	0.45	**0.6**
Ref. -2	-0.1	-0.05	**0**	-2.4	-3.0	-0.7	-0.2	0.35	**0.5**
Ref. -1 (in the dip)	2.3	2.35	2.4	**0**	-0.6	1.7	2.2	2.75	**2.9**

- Changing the reference year moves **all** coefficients by the same amount: same shape, different level.
- Ref. -1: the effect looks like 2.9 (the biased DiD), and all the years before training sit far from 0: the **reference year itself** is abnormal.
- Rule: the last normal year before the treatment, before any dip or anticipation. To find it, plot the **gap between the groups year by year** ($\bar{Y}_k^T - \bar{Y}_k^C$, which does not depend on a reference year), then check the event study: with a good reference, the pre-period $\hat{\beta}_k$ are close to 0.
- Check: two normal years (-4, -2) should give similar answers (0.6 vs. 0.5).

Standard errors: cluster by worker

- The same worker is observed in several years: their earnings are **correlated over time**, and their training status is set once.
- $n = I \times T$ rows, but only I independent workers: the usual (iid) standard errors are too small and the test rejects a true null too often (Bertrand et al. 2004, QJE).

$\Rightarrow$ **Cluster the standard errors by worker.**

- Pre-training coefficients close to 0 do not prove parallel trends: with noisy data, a real pre-trend can go unnoticed (Roth 2022, AER: Insights). (Why: see the appendix.)

When workers are treated at different dates

- With a single training date, the two-way FE estimate is the DiD of the four averages.
- With **staggered** dates, the two-way fixed effects estimate also compares late-treated workers with workers who were **already treated**. If effects grow over time, these comparisons are contaminated.

⇒ Use estimators that compare each cohort only with never-treated (or not-yet-treated) units: Sun and Abraham (2021, JoE), Callaway and Sant'Anna (2021, JoE). In R: `sunab()` in `fixest`.

In practice: three models, three lines of R

2×2: $Y_{it} = \beta_0 + \beta_1 \text{Trainee}_i + \beta_2 \text{Post}_t + \beta_3 (\text{Trainee}_i \times \text{Post}_t) + \varepsilon_{it}$

```
feols(earn ~ trainee * post, data = d, cluster = ~id)
```

Two-way FE: $Y_{it} = \gamma_i + \delta_t + \beta D_{it} + \varepsilon_{it}, \quad \text{with } D_{it} = \text{Trainee}_i \times \text{Post}_t$

```
feols(earn ~ D | id + year, data = d, cluster = ~id)
```

Event study (ref. -2): $Y_{it} = \gamma_i + \delta_t + \sum_{k \neq -2} \beta_k \mathbb{1}_{[t=k]} \times \text{Trainee}_i + \varepsilon_{it}$

```
feols(
   earn ~ i(year, trainee, ref = -2) | id + year,
   data = d, cluster = ~id
)
```

With `library(fixest)`. `earn`: Y_{it}, `trainee`: Trainee_i, `post`: Post_t, `id`: i, `year`: t. In the 2×2 output, β_3 is the line `trainee:post`.

6.5. Practice session

Practice session: difference-in-differences

Consider the following setting:

- From 2010, a program installs irrigation pumps in some villages.
- You are interested in measuring the effect of the program on rice yields (q/ha).
- The data: 200 villages, 2004–2015; half enter the program in 2010.

Objectives of the exercise

How do we check, estimate and display a DiD, and **when should we not trust it**?

Outline (about 60 min)

Parts 1–4	Simulated data, **the truth is known**: plot the trends, get the same DiD three ways, event study
Part 5	**DiD detective**: five provinces, the truth is hidden. One province per group: a number, a verdict, a suspect
Wrap-up	What the data can and cannot tell

Download the archive on Moodle. It contains the following files: the notebook, `did-lab-functions.R`, `project.Rproj`, `biblio.bib` and the folder `data/`, in the same directory. Required R Packages: `tidyverse`, `fixest`, `broom`, `here`.

7. Measuring causal impact: wrap-up

Measuring causal impact: the designs of Part A

Design	**Method**	**Key assumption**	**What to check**
Randomized trial	difference in means	randomization preserved	balance
Natural experiment, IV	difference in means; IV	relevance; exclusion	first stage; placebo groups
Regression discontinuity	comparison at the cutoff	no sorting at the cutoff	no piling up of units at the cutoff; characteristics
Difference-in-differences	change vs. change	parallel trends	a long pre-period

Three lessons from Part A

1. **The source of variation is the argument**: lotteries, cutoffs, comparison groups, instruments.
2. **Check what can be checked, argue the rest**: first stages, balance, pre-trends, piling up of units at a cutoff.
3. **One study is local**: look for what came next (comments, replications, replies).

Next: Part B, "how much is X worth?". A policy evaluation needs an **effect** (Part A) and a **value** (Part B).

8. Appendix

References I

Abadie, A., Athey, S., Imbens, G. W. and Wooldridge, J. M. (2022). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics* 138: 1–35, doi: 10.1093/qje/qjac038.

Abramitzky, R., Ager, P., Boustan, L., Cohen, E. and Hansen, C. W. (2023). The effect of immigration restrictions on local labor markets: Lessons from the 1920s border closure. *American Economic Journal: Applied Economics* 15: 164–191.

Angrist, J., Hull, P., Pathak, P. A. and Walters, C. R. (2024). Race and the mismeasure of school quality. *American Economic Review: Insights* 6: 20–37, doi: 10.1257/aeri.20220292.

Angrist, J. D. and Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics* 114: 533–575.

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives* 24: 3–30, doi: 10.1257/jep.24.2.3.

Ashenfelter, O. and Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics* 67: 648, doi: 10.2307/1924810.

Bakeev, M. (2023). Why not be data-driven? Historical arguments and their (ir)relevance to modern economics .

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: 22–53, doi: 10.1257/app.20130533.

References II

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy* 76: 169–217.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics* 119: 249–275, doi: 10.1162/003355304772839588.

Black, S. E., Denning, J. T. and Rothstein, J. (2023). Winners and losers? the effect of gaining and losing access to selective colleges on education and labor market outcomes. *American Economic Journal: Applied Economics* 15: 26–67, doi: 10.1257/app.20200137.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics* 225: 200–230, doi: 10.1016/j.jeconom.2020.12.001.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* 84: 772–793.

DellaVigna, S. and Malmendier, U. (2006). Paying not to go to the gym. *American Economic Review* 96: 694–719.

Deschênes, O. and Greenstone, M. (2007). The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather. *American Economic Review* 97: 354–385, doi: 10.1257/aer.97.1.354.

Deschênes, O. and Greenstone, M. (2011). Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US. *American Economic Journal: Applied Economics* 3: 152–185, doi: 10.1257/app.3.4.152.

References III

Donohue, J. J. and Wolfers, J. (2005). Uses and abuses of empirical evidence in the death penalty debate. *Stanford Law Review* 58: 791–845.

Ehrlich, I. (1975). The deterrent effect of capital punishment: A question of life and death. *American Economic Review* 65: 397–417.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review* 90: 980–994, doi: 10.1257/aer.90.4.980.

Garg, P. and Fetzer, T. (2025). Causal claims in economics.

Gertler, P. (2004). Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment. *American Economic Review* 94: 336–341, doi: 10.1257/0002828041302109.

Goldsmith-Pinkham, P. (2026). Tracking the credibility revolution across fields. doi: 10.3386/w35051.

Hastings, J. S. (2004). Vertical relationships and competition in retail gasoline markets: Empirical evidence from contract changes in Southern California. *American Economic Review* 94: 317–328, doi: 10.1257/000282804322970823.

Henrich, J., Heine, S. J. and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences* 33: 61–83.

Keane, M. P. (2010). A structural perspective on the experimentalist school. *Journal of Economic Perspectives* 24: 47–58, doi: 10.1257/jep.24.2.47.

References IV

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics* 114: 497–532.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review* 73: 31–43.

Leamer, E. E. (2010). Tantalus on the road to asymptopia. *Journal of Economic Perspectives* 24: 31–46, doi: 10.1257/jep.24.2.31.

L'Haridon, O. and Vieider, F. M. (2019). All over the map: A worldwide comparison of risk preferences. *Quantitative Economics* 10: 185–215, doi: 10.3982/qe898.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142: 698–714, doi: 10.1016/j.jeconom.2007.05.005.

Miguel, E. and Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72: 159–217, doi: 10.1111/j.1468-0262.2004.00481.x.

Nevo, A. and Whinston, M. D. (2010). Taking the dogma out of econometrics: Structural modeling and credible inference. *Journal of Economic Perspectives* 24: 69–82, doi: 10.1257/jep.24.2.69.

Nunn, N. (2008). The long-term effects of Africa's slave trades. *The Quarterly Journal of Economics* 123: 139–176.

Panhans, M. T. and Singleton, J. D. (2017). The empirical economist's toolkit. *History of Political Economy* 49: 127–157, doi: 10.1215/00182702-4166299.

References V

Romer, C. D. and Romer, D. H. (1989). Does monetary policy matter? A new test in the spirit of Friedman and Schwartz. In *NBER Macroeconomics Annual 1989*. MIT Press, *4*, 121–170.

Rosenbaum, P. (2017). *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press, doi: 10.4159/9780674982697.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights* 4: 305–322, doi: 10.1257/aeri.20210236.

Sala-i-Martin, X. (1997). I just ran two million regressions. *American Economic Review* 87: 178–183.

Scheiber, N. (2007). Freaks and geeks: How *Freakonomics* is ruining the dismal science. *The New Republic* : 27–31.

Sims, C. A. (2010). But economics is not an experimental science. *Journal of Economic Perspectives* 24: 59–68, doi: 10.1257/jep.24.2.59.

Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review* 66: 274–279.

Stock, J. H. (2010). The other transformation in econometric practice: Robust tools for inference. *Journal of Economic Perspectives* 24: 83–94, doi: 10.1257/jep.24.2.83.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* 225: 175–199, doi: 10.1016/j.jeconom.2020.09.006.

Urquiola, M. and Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review* 99: 179–215, doi: 10.1257/aer.99.1.179.

Quote from Leamer (1983)

Leamer (1983, AER)

The applied econometrician is like a farmer who notices that the yield is somewhat higher under trees where birds roost, and he uses this as evidence that bird droppings increase yields. However, when he presents this finding at the annual meeting of the American Ecological Association, another farmer in the audience objects that he used the same data but came up with the conclusion that moderate amounts of shade increase yields. A bright chap in the back of the room then observes that these two hypotheses are indistinguishable, given the available data. He mentions the phrase 'identification problem,' which, though no one knows quite what he means, is said with such authority that it is totally convincing.

Go back

Quote from Angrist and Pischke (2010)

Angrist and Pischke (2010, JEP)

Empirical evidence on any given causal effect is always local, derived from a particular time, place, and research design. Invocation of a superficially general structural framework does not make the underlying variation or setting more representative. Economic theory often suggests general principles, but extrapolation of causal effects to new settings is always speculative. Nevertheless, anyone who makes a living out of data analysis probably believes that heterogeneity is limited enough that the well-understood past can be informative about the future.

Go back

Same estimate, but which standard error?

	2×2 regression	**TWFE**
Estimate	**the same number**: $\hat{\beta}_3 = \hat{\beta}$	
In the residual ε_{it}	permanent differences between restaurants	only changes within a restaurant
Usual (iid) s.e.	too **large**	can be too **small**
s.e. clustered by restaurant	the same (up to a small-sample correction)	

- 2×2: the iid s.e. counts the permanent differences between restaurants as noise, although the DiD removes them.
- TWFE: with many periods, the iid s.e. ignores that shocks to a restaurant persist over time.

Why cluster by worker?



Errors ε_{it} of one worker, years -4 to 4; training in year 0 (made-up values).

- A worker's luck persists (health, job, employer): good and bad years come in **runs**. The same persistence shapes the Ashenfelter dip.
- Training is set once per worker: D_{it} is 0, then 1.

$\Rightarrow$ n rows (I workers $\times$ T years), but far fewer independent draws.

What the usual standard error misses

Let $\tilde{D}_{it}$ be D_{it} net of the worker and year fixed effects. Then

$$\hat{\beta} - \beta = \frac{\sum_i S_i}{\sum_i \sum_t \tilde{D}_{it}^2}, \qquad S_i = \sum_t \tilde{D}_{it}\, \varepsilon_{it} \quad \text{(worker } i\text{, all years)}$$

Workers are independent, so $\text{Var}(\hat{\beta})$ adds up the $\text{Var}(S_i)$, with

$$\text{Var}(S_i) = \underbrace{\sum_t \tilde{D}_{it}^2\, \text{Var}(\varepsilon_{it})}_{\text{iid and robust s.e. keep this...}} + \underbrace{\sum_{t \neq s} \tilde{D}_{it}\tilde{D}_{is}\, \text{Cov}(\varepsilon_{it}, \varepsilon_{is})}_{\text{...and drop this}}$$

- Two years before training (or two after): $\tilde{D}_{it}\tilde{D}_{is} > 0$, and close years have correlated errors.

$\Rightarrow$ The dropped term is **positive**: the usual s.e. is too small.

The clustered standard error

Robust: one worker-year at a time

$$\frac{\sum_i \sum_t \left(\tilde{D}_{it}\,\hat{\varepsilon}_{it}\right)^2}{\left(\sum_i \sum_t \tilde{D}_{it}^2\right)^2}$$

Square each year, then add:
the products $\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}$ are dropped.

Clustered by worker

$$\frac{\sum_i \left(\sum_t \tilde{D}_{it}\,\hat{\varepsilon}_{it}\right)^2}{\left(\sum_i \sum_t \tilde{D}_{it}^2\right)^2}$$

Add the years of a worker, then square:
the products are kept, whatever their pattern.

- No model of the correlation is needed, only that workers are independent.
- It needs many workers.

At which level should we cluster?

- At the level where the treatment is assigned (Abadie et al. 2022, QJE): training is decided worker by worker $\Rightarrow$ cluster by worker.
- If training were decided firm by firm: cluster by **firm**.
- With few clusters, the clustered s.e. is unreliable. In Card and Krueger (1994, AER), the minimum wage is set by state: only 2 clusters.

10. Backup slides

Backup: two details from Ashenfelter and Card (1985)

1. The comparison group contains some trainees

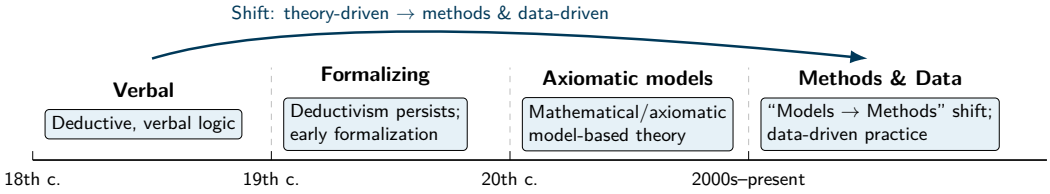
- The comparison group is a sample of the **whole population**, trainees included: the estimated gap is too small by a factor $(1 - p)$, p the share of participants.

2. A DiD with a base year before the dip is not the paper's final estimator

- The authors model earnings (permanent component, year effects, serially correlated shocks) and participation.
- They **predict** what trainees would have earned without training.

Backup: a shift from theory-driven to data-driven discipline

From Bakeev (2023), we understand that our discipline is not fixed, it is evolving:



- Since the 1970s, economics has transitioned toward empirical work (Panhans and Singleton 2017, Hist. Political Econ.)