IV: When the Coin Is Bent

Gov 51: Data Analysis and Politics

Scott Cunningham

Harvard University

Week 12
April 16, 2026

Tuesday: IV works when the coin is fair and strong

- Selection bias breaks OLS even at the population level
- IV recovers β_1 by exploiting exogenous variation in D
- Wald = RF $\div$ FS = β_1 (under the three conditions)
- 2SLS = Wald with controls — same logic
- AJR: settler mortality $\rightarrow$ institutions $\rightarrow$ income

Today: what happens when the instrument barely moves D? The denominator of the Wald ratio approaches zero. Everything breaks.

Today's roadmap

1. **The bent coin:** Angrist-Krueger (1991) and what "weak" means
2. **How weak instruments break 2SLS:** bias and variance
3. **Testing instrument strength:** the F-statistic, Stock-Yogo, Olea-Pflüger
4. **Anderson-Rubin CIs:** a confidence interval that doesn't require a strong FS
5. **LATE:** what IV actually estimates — Angrist, Imbens, and the compliers
6. **Finding the coin flip:** natural experiments in the wild
7. **Aizer-Doyle (2015):** juvenile incarceration and judge leniency
8. **Albouy (2012):** how bad data can fake a strong instrument

A strong first stage is not optional. It is load-bearing for everything else IV does.

Part I: The Bent Coin
Angrist and Krueger (1991)

IV works by finding a coin flip — Z is randomly assigned and uses that randomness to move D

1. Find a variable Z that is as-good-as-randomly assigned
2. Z shifts D for *some* units (called compliers, discussed later)
3. Compare outcomes across values of Z: Wald $= \Delta Y \,/\, \Delta D$

Z must **cause** D, not merely correlate with it. A correlation is not a mechanism — confidence comes from institutional knowledge, not a coefficient.

Three conditions must hold — and the coin must truly be the source of variation in D

- **Exclusion:** Z affects Y *only* through D — no back door
- **Independence:** $Z \perp\!\!\!\perp U$ — the coin is fair
- **Relevance:** $\text{Cov}(Z, D) \neq 0$ — the coin actually moves D

Finding a correlation between Z and D is not sufficient.

You must be confident that Z is the *mechanism* driving some units into and out of treatment — not just associated with it.

Instrument strength is a see-saw — small data needs a large first stage, weak first stage needs large data

- Power $= f(\text{ effect size, } N)$ — two levers, one see-saw
- Small $N \Rightarrow$ need large $\text{Cov}(Z, D)$ to reject $H_0 \colon \pi = 0$
- Small $\text{Cov}(Z, D) \Rightarrow$ need large N to reject $H_0 \colon \pi = 0$

A weak instrument does not merely fail to help.

As $\hat{\pi}_{\text{FS}} \to 0$, the Wald denominator shrinks and 2SLS bias converges to OLS bias — the estimator **amplifies** the endogeneity it was meant to fix.

Angrist-Krueger (1991) made weak instruments a household name — it ran in the *AER*

The American Economic Review is the flagship journal of economics — the *Science* or *Nature* of the discipline. High-profile placement means high-profile scrutiny.

AK did not invent the weak instrument problem. They made it famous by putting a weak instrument in a paper the entire profession read.

Quarter of birth → schooling → wages. High prestige. Hidden flaw.

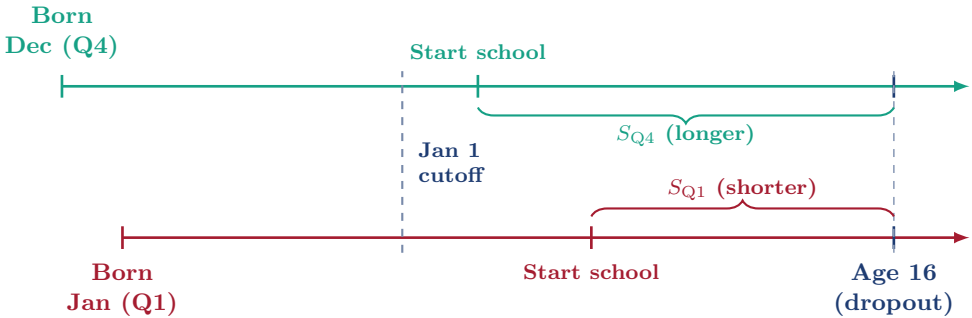
Angrist and Krueger (1991): your birthday determines how much school you're forced to attend

The instrument: quarter of birth

- Compulsory schooling laws: you can drop out at age 16
- Born in Q1 (January): start school earlier, reach 16 with *fewer* years completed
- Born in Q4 (October): start later, must accumulate more years to reach 16
- Your birthday is quasi-random — no one chooses when to be born

$$Z = \text{quarter of birth} \quad \to \quad D = \text{years of education} \quad \to \quad Y = \log \text{wages}$$

Q4-born children accumulate more compulsory schooling before the legal dropout age



Jan 1 cutoff forces Q1-born kids to wait an extra year to enroll
$\Rightarrow S_{Q4} > S_{Q1}$. The law, not ability, creates the difference.

But the coin was bent: quarter of birth barely moves education

- Q1 vs. Q4 difference in schooling: only ≈ 0.1 years
- The Wald denominator — $\text{Cov}(D, Z)$ — is near zero
- AK added 180 instruments (quarters $\times$ states $\times$ years) to try to compensate

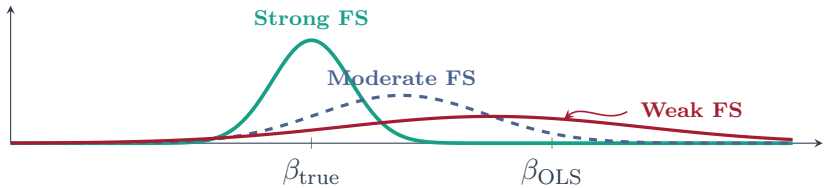
The bent coin: $\text{Cov}(D, Z) \approx 0$ so Wald $= \dfrac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)}$ is dividing by near-zero.

More instruments raised the denominator mechanically — but each was weak, and overfitting Stage 1 made the problem worse, not better.

Part II: How Weak Instruments Break 2SLS Bias and Variance

As the first stage weakens, the IV distribution slides toward OLS bias

As $\text{Cov}(D, Z) \to 0$, the Wald denominator shrinks — bias *and* variance both worsen:



As the first stage collapses: IV bias slides toward OLS bias and variance explodes simultaneously — two problems at once.

The structural equation contains everything you didn't measure

$$Y_i = \beta_0 + \beta_1 D_i + \underbrace{u_i}_{\text{everything else}}$$

u_i = ability, background, neighborhood, luck

every omitted cause of Y that you did not put in the model

OLS recovers β_1 only if $\text{Cov}(u, D) = 0$.
In social science, that is almost never true.

The bias is a regression coefficient — one you can never run

$$\hat{\beta}_1^{\text{OLS}} = \beta_1 + \underbrace{\frac{\text{Cov}(u,\, D)}{\text{Var}(D)}}_{\text{the bias}}$$

$\frac{\text{Cov}(u,\, D)}{\text{Var}(D)}$ is the slope from regressing u on D

You cannot run that regression — u is unobserved by definition.
If $\text{Cov}(u, D) \neq 0$: the bias is real, it has a sign, and more data never fixes it.

The bias formula makes the drift precise

$$E\left[\hat{\beta}_{IV}\right] \approx \beta_1 + \underbrace{\frac{\mathrm{Cov}(D,\, u)}{\mathrm{Var}(D)}}_{\text{OLS bias}} \cdot \frac{1}{F+1}$$

First stage F	IV bias relative to OLS bias	Verdict
$F = 100$	$1/101 \approx 0\%$	essentially unbiased
$F = 10$	$1/11 \approx 9\%$	acceptable
$F = 3$	$1/4 = 25\%$	dangerous
$F = 1$	$1/2 = 50\%$	cure = disease

When $F \approx 1$: IV bias $\approx$ OLS bias. You fixed nothing.

Variance explodes because the Wald denominator approaches zero

Let $\rho_{ZD} = \text{Corr}(Z, D)$: how strongly the instrument moves the treatment.

$$\frac{\text{Var}(\hat{\beta}_{IV})}{\text{Var}(\hat{\beta}_{OLS})} \quad = \quad \frac{1}{\rho_{ZD}^2}$$

Correlation ρ_{ZD}	Variance ratio (IV vs. OLS)	SE ratio
0.7	2×	1.4×
0.3	11×	3.3×
0.1	100×	10×

Wald = $\text{Cov}(Y, Z) \,/\, \text{Cov}(D, Z)$. Denominator ≈ 0 $\Rightarrow$ ratio jumps wildly across samples. That *is* variance.

IV is consistent — but "consistent" is not the same as "better"

Consistent: $\hat{\beta}_{IV} \xrightarrow{p} \beta_1$ as $n \to \infty$

- OLS is biased *forever* — more data never helps
- IV is biased in finite samples, but the bias shrinks as n grows

The catch: convergence is slow when ρ_{ZD} is small

- Normal approximation requires $n \cdot \rho_{ZD}^2$ to be large
- Weak instrument: need $n = 10{,}000$ to do what $n = 100$ does with a strong one

Being consistent does not mean IV always beats OLS in a finite sample.
With a weak enough instrument, IV can be *worse*.

MSE = Var + Bias^2 — the right scorecard for comparing OLS and IV

$$\text{MSE}(\hat{\beta}) \quad = \quad \text{Var}(\hat{\beta}) \quad + \quad \text{Bias}(\hat{\beta})^2$$

Total error = how spread out your estimates are + how far they drift from the truth.

	OLS	**IV (weak FS)**
Variance	Low	Very high
Bias^2	Permanent, never zero	Shrinks as $n \to \infty$
MSE verdict	Moderate, stable	Can exceed OLS

A weak instrument inflates variance so much that MSE(IV) > MSE(OLS) — even though OLS is permanently wrong.

Part III: Testing Instrument Strength
The F-Statistic

A strong instrument moves the treatment; a weak one barely does

Strong instrument	Weak instrument
Z explains much of D	Z explains almost none of D
$\text{Cov}(D, Z)$ is large	$\text{Cov}(D, Z) \approx 0$
Wald denominator stable	Wald: dividing by near-zero
Estimates precise, near β_1	Estimates explode sample to sample

We need a number to summarize "how strong" the first stage is. That number is the **F-statistic**.

The F-statistic asks: does adding the instrument improve the fit of the first stage?

Two regressions for the first stage:

- **Restricted:** $D \sim$ controls (no instrument) $\rightarrow SSR_R$
- **Unrestricted:** $D \sim Z +$ controls (with instrument) $\rightarrow SSR_U$

If Z is relevant, adding it should reduce the residuals: $SSR_U < SSR_R$.

$$F = \frac{(SSR_R - SSR_U) \,/\, q}{SSR_U \,/\, (n - k - 1)}$$

If the instrument explains a lot of D, $SSR_U \ll SSR_R$ and F is large.

Computing F by hand with Card (1995)

Card data: $n = 3{,}010$, $q = 1$ instrument (nearc4), $k = 6$ controls

Model	SSR
Without nearc4	$SSR_R = 11{,}395$
With nearc4	$SSR_U = 11{,}332$

$$F = \frac{(11{,}395 - 11{,}332) \,/\, 1}{11{,}332 \,/\, (3{,}010 - 6 - 1)} = \frac{63}{3.77} \approx 16.7$$

R gives the same answer: $t^2 = F$ when there is one instrument

```
fs <- lm(educ ~ nearc4 + exper + expersq +
                black + south + smsa,
         data = card)

summary(fs)   # t-stat on nearc4 = 4.09
              # F = t^2 = 4.09^2 = 16.7
```

With one instrument, $F = t^2$. The t-test on the instrument in the first stage *is* the F-test.

How high must F be? — $F > 10$ is the threshold, but use the effective F for real work

With one instrument: $F = t^2$. The t-test on the instrument *is* the F-test.

With multiple instruments: F tests them *jointly* — individual t-stats are not enough.

Rule	Source	Limitation
$F > 10$	Stock-Yogo (2005)	Assumes homoskedastic errors
Effective $F > 10$	Olea-Pflüger (2013)	Robust to heteroskedasticity

In R: `iv_robust()` reports the Olea-Pflüger effective F automatically. Use it. Do not rely on the Stock-Yogo $F = 10$ rule for real research.

More weak instruments overfits Stage 1 — noise flows into Stage 2

- With L instruments, Stage 1 has L regressors
- Each weak instrument explains a tiny slice of D
- Together, they overfit: $\hat{D}$ starts absorbing *noise*
- Stage 2 regresses Y on that noise $\rightarrow$ garbage estimates

You already know this problem. Adding useless predictors to a regression causes overfitting. This is the same thing — applied to Stage 1.

One strong instrument beats ten weak ones

$$F = 50,\ 1 \text{ instrument} \quad \gg \quad F = 3,\ 10 \text{ instruments}$$

- More instruments exacerbates bias when each is weak
- The effective F does not add up across weak instruments
- AK (1991): 180 instruments, each nearly useless

Better instruments = more credibility. More instruments $\neq$ more credibility.

Wait — OVB predicts IV should be *below* OLS. Card found the opposite.

$$\hat{\beta}^{\text{OLS}} = \beta_1 + \underbrace{\frac{\text{Cov}(\text{ability}, D)}{\text{Var}(D)} \cdot \lambda}_{> 0 \text{ (ability bias)}}$$

- Ability ↑ ⇒ more schooling *(Cov(ability, D) > 0)*
- Ability ↑ ⇒ higher wages *(λ > 0)*
- OVB formula predicts: IV < OLS (IV removes the upward bias)

Prediction from OVB: Card's IV should come in *below* his OLS estimate.

Card (1995) found the opposite — 2SLS > OLS — and it surprised labor economists

OVB formula predicts ability bias pushes OLS *up* — so IV should come in *below* OLS.

OLS = 0.074 2SLS = 0.132 2SLS is almost *twice* as large

Lower ability ⇒ higher wages? That doesn't make sense.
Something else is going on.

Card's explanation: college doesn't help everyone the same way

- The OVB story assumes one kind of student with one kind of return
- But what if college helps *different students differently?*
- Students near a college who couldn't afford to go otherwise — credit-constrained, not low-ability — may have *unusually high* returns when they finally do go

College proximity selects students who attend *only* because a college was nearby. If that group has higher returns than average, the effect IV measures exceeds the population average — and IV > OLS makes perfect sense.

What IV actually estimates depends on who the instrument moves. That is where we go next.

Part IV: Anderson-Rubin Confidence Intervals Valid Even When the First Stage Is Weak

Standard IV confidence intervals break down when the instrument is weak

The standard 2SLS CI: $\hat{\beta}_{IV} \pm 1.96 \times \widehat{SE}$

- The SE formula assumes the first stage is strong
- When F is small: SE underestimates true uncertainty
- The CI looks tight — but it is centered on an unreliable estimate

Weak instrument $\Rightarrow$ Wald = RF/FS is dividing by near-zero.
The CI inherits that instability. More data does not fix it.

Anderson-Rubin: invert a test instead of scaling an estimate

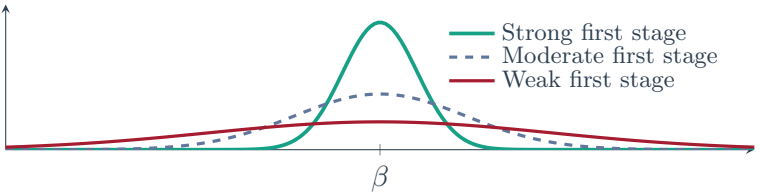
The idea (plain statistics):

- For each candidate value β_0, form the residual $\tilde{e}_i = Y_i - \beta_0 D_i$
- If $\beta = \beta_0$ truly, then $\tilde{e}_i$ should be *uncorrelated with* Z
- Test $H_0 \colon \mathrm{Cov}(\tilde{e}, Z) = 0$ using a t-test on Z
- Collect all β_0 that *pass* the test (cannot be rejected) $\rightarrow$ the AR confidence set

The AR CI is the set of β_0 values *consistent with the data* at the 95% level.
No division by the first stage. No weak-instrument problem in the CI construction.

When the first stage is weak, the sampling distribution of $\hat{\beta}_{\text{IV}}$ spreads and AR is valid

$\hat{\beta}_{\text{IV}} = \text{RF}/\text{FS}$ — as $\text{FS} \to 0$, dividing by near-zero makes the distribution flat and uninformative:



$\hat{\beta}_{\text{IV}}$ (centered at true β)

AR never divides by the first stage — it tests whether Z predicts $Y - \beta_0 D$ in plain OLS. Valid at any F.

Example: Card (1995) standard CI vs. AR CI

	Standard 2SLS CI	Anderson-Rubin CI
Point estimate	0.132	—
95% CI	[0.04, 0.23]	[0.04, 0.26]

- AR CI is wider on the upper end — it is more honest about uncertainty
- Difference is small because $F \approx 17$ (strong-ish instrument)
- With $F \approx 3$: the AR CI can be *much* wider, or even unbounded

When $F > 20$: standard and AR CIs are nearly identical. When $F \approx 10$: use AR.

The AR CI in R: grid inversion

For one instrument: AR test = t-test on Z in OLS of $\tilde{e}_i = Y_i - \beta_0 D_i$ on Z

```r
library(wooldridge); data(card)

# Grid-inversion: collect beta0 values we cannot reject
beta_grid <- seq(-0.1, 0.5, by = 0.001)
pvals <- sapply(beta_grid, function(b0) {
  e_tilde <- card$lwage - b0 * card$educ
  m <- lm(e_tilde ~ nearc4 + exper + expersq + black + south + smsa,
          data = card)
  coef(summary(m))["nearc4", 4]   # p-value on Z (this IS the AR test)
})
ar_ci <- range(beta_grid[pvals > 0.05])   # [0.039, 0.261]
```

The p-value on `nearc4` IS the AR test p-value. No special package required.

Part V: LATE
What IV Actually Estimates

Heckman (1990): unless Z can move *everyone*, IV cannot identify the average treatment effect

- IV exploits variation in Z to recover a causal effect
- But Z never moves everyone — some always take the treatment, some never do
- If the instrument affects only *some* people, it can't average over the full population

Heckman called this "identification at infinity" — you'd need an instrument powerful enough to shift the *entire* population. That instrument almost never exists.
So what *does* IV identify?

Angrist (1990): the Vietnam draft lottery was an actual coin flip — not a proxy, a randomization

- His Princeton dissertation and job market paper
- **The instrument:** draft lottery random sequence numbers (RSNs)
- **Data:** Selective Service records + SSA earnings histories
- **First stage:** lottery raised military service by ≈ 15 percentage points

Unlike most instruments, this *was* the randomization.
The government literally drew numbers from a bowl.
No selection. No earnings correlation. Just a lottery.

Imbens and Angrist (1994) answered Heckman: IV recovers an average effect — but only for compliers

How they met (c. 1990 at Harvard):

- Angrist joins the Harvard faculty; Imbens arrives from Erasmus Rotterdam via Brown PhD
- Imbens got the job by phone after the market cleared — thought he was overplaced
- Both knew Heckman's challenge. Both knew the draft lottery paper. What reconciles them?

The paper: *Imbens & Angrist*, Econometrica 1994 (listed out of alphabetical order: I before A)

IV *does* recover an average effect — but only for the units whose treatment status is actually moved by the instrument.

The population divides into four types — IV finds only one of them

Type	Responds to the lottery?	IV finds their effect?
Always-takers	Enroll whether drafted or not	No
Never-takers	Never enroll, draft or not	No
Compliers	**Enroll iff drafted**	**Yes — LATE**
Defiers	Enroll iff *not* drafted	Must not exist

All four types have treatment effects.
IV identifies only the compliers.

Always-takers: patriots and warriors — the instrument is irrelevant to them

- Enroll in the military *regardless* of their lottery number
- Not drafted? Volunteer anyway.
- $D_i(Z = 1) = D_i(Z = 0) = 1$

They have a treatment effect.
The instrument has no leverage over their decision.
IV cannot find their effect.

Never-takers: conscientious objectors — the instrument is also irrelevant to them

- Refuse to serve *regardless* of their lottery number
- Drafted? Flee to Canada. Go to prison. Claim exemption.
- $D_i(Z = 1) = D_i(Z = 0) = 0$

They also have a treatment effect.
The instrument has no leverage over their decision either.
IV cannot find their effect.

Compliers: they serve *if and only if* drafted — LATE is the average effect for this group

- Value obedience: when the lottery draws their number, they go
- If not drafted, they don't volunteer on their own
- $D_i(Z = 1) = 1, \quad D_i(Z = 0) = 0$

LATE = average treatment effect for compliers.
This is what the Wald estimator and 2SLS identify.
"Local" means local to *this* group — not the full population.

Defiers: they do the opposite of their assignment — and monotonicity rules them out

- Value disobedience: if drafted, they refuse; if not drafted, they volunteer
- $D_i(Z = 1) = 0, \quad D_i(Z = 0) = 1$
- Not impossible (protest, counterculture) — but assumed negligibly rare

Monotonicity assumption: no defiers (or so few as to be irrelevant). Without it, compliers and defiers cancel in the Wald denominator and LATE is not identified.

The four assumptions for LATE — monotonicity is the new one

1. **Independence:** $Z \perp\!\!\!\perp (Y_0, Y_1, D_0, D_1)$ — *the coin is fair*
2. **Exclusion:** $Z \to Y$ only through D — *no back door*
3. **Relevance:** $\Pr(D = 1 \mid Z = 1) \neq \Pr(D = 1 \mid Z = 0)$ — *coin moves D*
4. **Monotonicity:** $D_i(Z = 1) \geq D_i(Z = 0)$ for all i — *no defiers*

Monotonicity does not require the instrument to move everyone the same amount.
It only requires that no one does the *opposite* of their assignment.

Part VI: Finding the Coin Flip
Natural Experiments in the Wild

IV is about finding randomization that already happened in the world

The wrong question: "Which variable correlates with D?"

The right question: "Where did nature, policy, or chance create *random* variation in D?"

A good instrument is a **coin flip you didn't design** — but that you found.

- The coin flip creates exogenous variation in D
- You cannot design it after the fact
- You must know enough about the *world* to recognize it when you see it
- IV skill is less about econometrics and more about institutional knowledge

Natural experiments appear everywhere — if you know what to look for

Instrument	Treatment	The coin flip
Judge leniency	Incarceration	Random case assignment
Rainfall growth	Economic growth	Weather variation
College proximity	Years of schooling	Distance at birth
Quarter of birth	School starting age	Birthday
Draft lottery number	Military service	Random draw
Settler mortality	Institutions	Disease environment

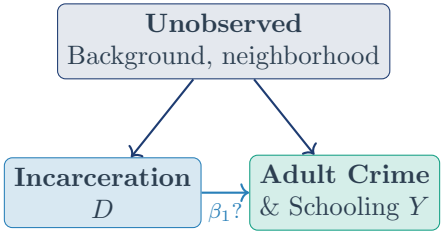
Each coin flip passes the strangeness test: the reduced form is puzzling *until* you know the treatment story.

Part VII: Aizer and Doyle (2015)
The Causal Effect of Juvenile Incarceration

A 13-year-old arrives in Chicago court — who decides their fate?

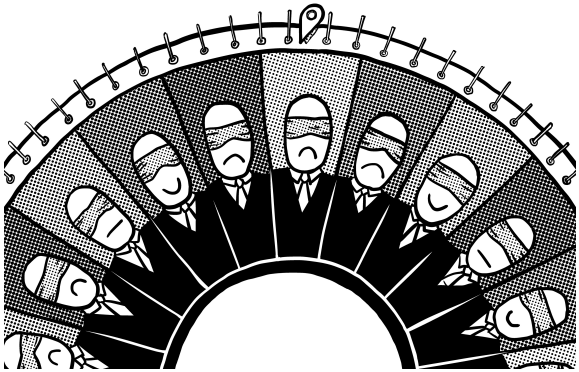
Does juvenile incarceration cause worse adult outcomes?

- US has the highest juvenile corrections rate in the world — 70,000 detainees in 2010
- Kids who get incarcerated are not random: prior record, home situation, neighborhood
- OLS: incarceration → adult crime +41pp (upward bias from confounders)



OLS mixes β_1 with confounder paths

Judge assignment is the coin flip



Judges vary in leniency. Cases are randomly assigned. Strict judge → more likely incarcerated.

Incarceration increases adult crime — but OLS overstates it

TABLE V

JUVENILE INCARCERATION AND ADULT CRIME

	Dependent variable: entered adult prison by age 25						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Full CPS sample			Juvenile court sample			
	OLS	OLS	Inverse propensity score weighting	OLS	OLS	2SLS	2SLS
Juvenile incarceration	0.407	0.350	0.219	0.200	0.155	0.260	0.234
	(0.0082)	(0.0064)	(0.013)	(0.0072)	(0.0073)	(0.073)	(0.076)
Demographic controls	No	Yes	Yes	No	Yes	No	Yes
Court controls	N/A	N/A	N/A	No	Yes	No	Yes
Observations	440797	440797	420033	37692			
Mean of dependent variable	0.067	0.067	0.057	0.327			

	OLS	2SLS
Adult incarceration	+41pp	+23pp

Incarceration reduces high school graduation — OLS again overstates

TABLE IV

JUVENILE INCARCERATION AND HIGH SCHOOL GRADUATION

	Dependent variable: graduated high school						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Full CPS sample			Juvenile court sample			
	OLS	OLS	Inverse propensity score weighting	OLS	OLS	2SLS	2SLS
Juvenile incarceration	−0.389	−0.292	−0.391	−0.088	−0.073	−0.108	−0.125
	(0.0066)	(0.0065)	(0.0055)	(0.0043)	(0.0041)	(0.044)	(0.043)
Demographic controls	No	Yes	Yes	No	Yes	No	Yes
Court controls	N/A	N/A	N/A	No	Yes	No	Yes
Observations	440,797	440,797	420,033	37,692			
Mean of dependent variable	0.428	0.428	0.433	0.099			

	OLS	2SLS
High school graduation	−39pp	−13pp

Compliers: the marginal kids whose fate the judge's coin flip changed

Type	Who they are
Always-incarcerated	Jailed regardless of which judge they drew
Never-incarcerated	Released regardless of which judge they drew
Compliers	**Fate tipped by the judge — the marginal kid**

LATE = effect of incarceration for compliers only.

2SLS is smaller than OLS because compliers are the marginal, less severe cases

	OLS	2SLS
Adult incarceration	+41pp	+23pp
High school graduation	−39pp	−13pp

2SLS is smaller because the compliers — kids on the margin — are less severe cases than those who would always be incarcerated. The judge's coin flip identifies the marginal kid, not the hardened offender.

Part VIII: A Skeptic Challenges AJR
The Albouy (2012) Critique

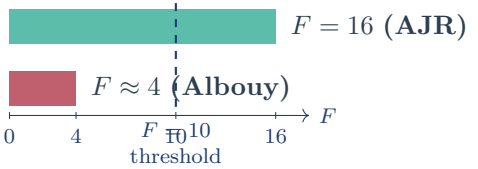
AJR (2001): settler mortality → institutions → income — $F = 16$, result significant

- Z = log settler mortality, 1817–1848
- First stage $F = 16$ — comfortably above the Stock-Yogo threshold
- 2SLS coefficient = 0.917, statistically significant
- Institutions explain a large share of income differences across former colonies

Strong instrument, strong result. The paper won the 2024 Nobel Prize.

Albouy (2012): mortality rates were assigned to the wrong countries

- Coding errors in $\approx 1/3$ of observations
- Mortality rates copied across countries within the same region
- With corrected data: F drops from 16 to ≈ 4
- Results lose statistical significance



Lesson: a high F from bad data is noise, not instrument strength

The F-test measures whether Z predicts D.
It *cannot* tell you whether Z predicts D for the right reasons.

- A mismeasured Z can correlate with D spuriously
- A high F on noisy data says: "this noise correlates with treatment"
- That is not the same as: "this instrument is valid"
- The exclusion restriction is still your job to argue, not the F-test's

The Nobel Committee (2024) sided with AJR. But Albouy's critique made the profession far more careful about instrument measurement quality.

The IV checklist

Condition	How to check	If it fails
Relevance	$F > 10$; Olea-Pflüger	Weak instrument: bias problem
Exclusion	Theory only	Abandon or redesign
Independence	Natural experiment	Abandon or redesign
LATE scope	Describe compliers	Report who it applies to

Only relevance is testable. The other three require knowing the world.

A strong first stage is not optional. Without it, IV is just OLS in disguise.