Research Methodology for Public Policy

Lecture 9 — Experiments and Causation

Dr. Colin Kuehl

2026-07-21

Northern Illinois University

SCHOOL OF PUBLIC POLICY
CHIANG MAI UNIVERSITY

Welcome Back & Agenda

Today's Agenda

- Recap & questions from Lecture 9
- More Regression
  - Interpreting Binary and Categorical
  - Useful Transformations
  - Regression Assumptions
- Experiments
  - Why experiments are political science's best tool for causal identification
  - Random assignment vs. random sampling — a distinction worth getting exactly right
  - Four types of experiments,
  - Internal, external, and construct validity — the three-way trade-off
  - Advantages, disadvantages, and what can still go wrong
  - Power, pre-registration, and research integrity
- Midterms Returned

Why Regression Is So Useful (and Where It Breaks)

Where Regression Shines, and Where It Breaks

Strengths

- **Generalizable** beyond the cases observed, if the sample represents the population well
- **Precise**: a number plus an uncertainty estimate, not just "yes" or "no"
- Easy to **control for other variables** in one model
- Supports **iteration**: add a variable, refit, learn something

Limitations

- **Measurement**: garbage in, garbage out
- **Average effects** can mask subgroup variation
- **Unrealistic causal claims** from correlational designs
- **Kitchen-sink regressions** risk data mining

Interpreting Dummy and Categorical Predictors

- Not every X is continuous — a **dummy variable** codes a category as 0/1 (e.g., democracy vs. not). Its coefficient is the average difference in Y between the two groups, holding other variables constant. More in Practicum

Dummy

- A dummy variable is equivalant to a difference in means between the two groups.

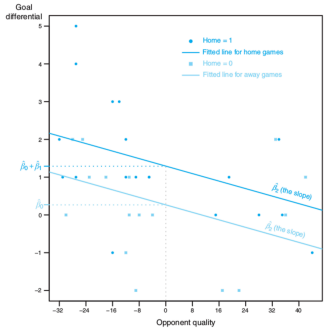


FIGURE 6.7: Fitted Values for Model with Dummy Variable and Control Variable: Manchester City Example

Transforming Variables

Why Transform Variables?

Transformations serve two purposes:

1. **Fix model problems** — address non-linearity, heteroskedasticity, or skew
2. **Aid interpretation** — make coefficients more substantively meaningful

Transformation Types

Transformation	Typical use case
Natural log $\ln(X)$	Right-skewed, positive variables; diminishing returns
Z-score (standardise)	Comparing effects across different-scale predictors
Square root $\sqrt{X}$	Count data with moderate skew
Squared term X^2	U-shaped or inverted-U relationships
First difference ΔX_t	Non-stationary time series

Log Transformation

When to use: X is strictly positive and right-skewed (GDP, population, income).

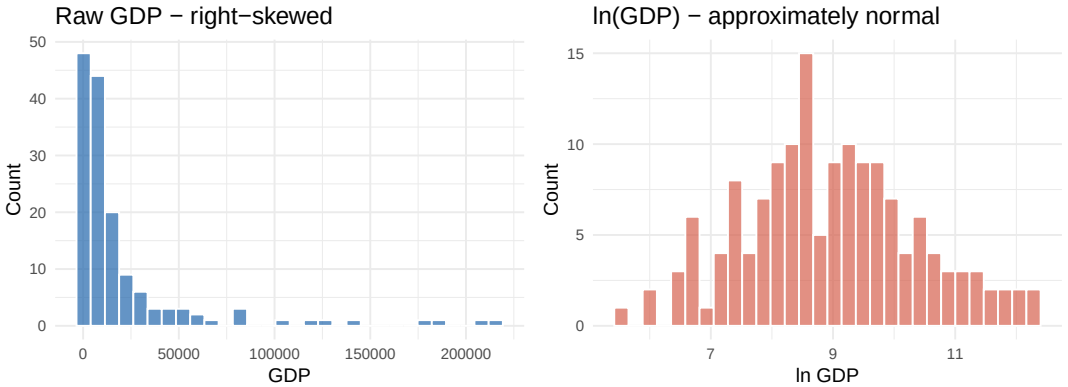


Figure 1

Interpreting Log Models

Interpretation changes depending on **which side** of the equation is logged.

Table 1

Model	Specification	Interpretation
Level–Level	$Y = \beta X$	1-unit $\uparrow$ in $X \rightarrow \beta$-unit change in Y
Log–Level	$\ln Y = \beta X$	1-unit $\uparrow$ in $X \rightarrow (e^{\beta} - 1) \times 100\%$ change in Y
Level–Log	$Y = \beta \ln X$	1% $\uparrow$ in $X \rightarrow \beta/100$-unit change in Y
Log–Log	$\ln Y = \beta \ln X$	1% $\uparrow$ in $X \rightarrow \beta\%$ change in Y (elasticity)

Log–log is common in economics (demand elasticity). Log–level and level–log

Z-Score Standardisation

Formula: $\tilde{X} = \dfrac{X - \bar{X}}{s_X}$

The standardised variable has **mean 0** and **standard deviation 1**.

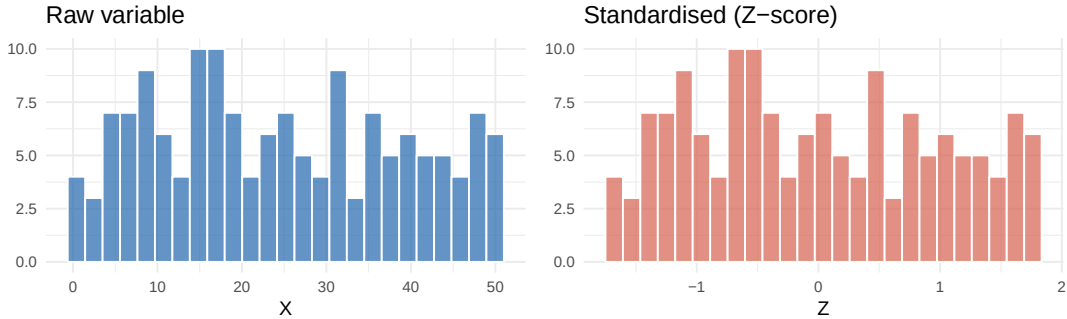


Figure 2

When to Use Z-Scores

Primary purpose: Make regression coefficients **directly comparable** when predictors are on different scales.

Table 2

	Raw	Standardised
(Intercept)	-4.530***	-0.025
	(0.556)	(0.082)
income	0.000***	
	(0.000)	
edu_yr	0.229***	
	(0.038)	
scale(income)		0.392***
		(0.084)
scale(edu_yr)		0.505***
		(0.084)

When to standardise

- Comparing substantive effect sizes
- Predictors measured in incompatible units
- Presenting results to non-technical audiences

When *not* to standardise

- When the raw unit is substantively meaningful (years, %)
- Binary / dummy variables (don't standardise)
- When comparing coefficients

Transformation Checklist

Before running your final model, ask:

1. **Is my outcome or key predictor right-skewed?** $\rightarrow$ Consider $\ln(Y)$ or $\ln(X)$; check residual normality after.

Transformation Checklist

Before running your final model, ask:

1. **Is my outcome or key predictor right-skewed?** $\rightarrow$ Consider $\ln(Y)$ or $\ln(X)$; check residual normality after.
2. **Are my predictors on wildly different scales?** $\rightarrow$ Standardise for presentation; keep raw for policy-relevant interpretation. OR just change units (ie thousands of gdp)

Transformation Checklist

Before running your final model, ask:

1. **Is my outcome or key predictor right-skewed?** $\rightarrow$ Consider $\ln(Y)$ or $\ln(X)$; check residual normality after.
2. **Are my predictors on wildly different scales?** $\rightarrow$ Standardise for presentation; keep raw for policy-relevant interpretation. OR just change units (ie thousands of gdp)
3. **Does the residual vs. fitted plot show a funnel?** $\rightarrow$ Log-transforming Y or using robust SE.

Transformation Checklist

Before running your final model, ask:

1. **Is my outcome or key predictor right-skewed?** $\rightarrow$ Consider $\ln(Y)$ or $\ln(X)$; check residual normality after.
2. **Are my predictors on wildly different scales?** $\rightarrow$ Standardise for presentation; keep raw for policy-relevant interpretation. OR just change units (ie thousands of gdp)
3. **Does the residual vs. fitted plot show a funnel?** $\rightarrow$ Log-transforming Y or using robust SE.
4. **Does the relationship look curved?** $\rightarrow$ Add X^2, or transform to $\ln(X)$.

Transformation Checklist

Before running your final model, ask:

1. **Is my outcome or key predictor right-skewed?** $\rightarrow$ Consider $\ln(Y)$ or $\ln(X)$; check residual normality after.
2. **Are my predictors on wildly different scales?** $\rightarrow$ Standardise for presentation; keep raw for policy-relevant interpretation. OR just change units (ie thousands of gdp)
3. **Does the residual vs. fitted plot show a funnel?** $\rightarrow$ Log-transforming Y or using robust SE.
4. **Does the relationship look curved?** $\rightarrow$ Add X^2, or transform to $\ln(X)$.
5. **Is VIF $>$ 5 for any variable?** $\rightarrow$ Check whether logged or standardised versions reduce collinearity.

Summary

Table 3

Problem	What it affects	Coefficients biased?	Key diagnostic	Primary fix
Multicollinearity	Standard errors (inflated)	No	VIF $>$ 5–10	Drop / composite variable
Heteroskedasticity	Standard errors (wrong sign)	No	Residual plot; Breusch-Pagan	HC3 robust SE
Autocorrelation	Standard errors (underestimated)	No	ACF plot; Durbin-Watson	HAC / clustered SE

All three problems corrupt **inference** (hypothesis tests, confidence intervals), not estimation. Always report and defend your diagnostic choices.

Regression Assumptions and (how to fix when you break them)

The Classical Linear Regression Model

OLS produces unbiased, efficient estimates **only when its assumptions hold.**

Table 4

#	Assumption	If Violated
A1	Linearity	Biased estimates
A2	Random Sampling	Non-representative results
A3	No Perfect Multicollinearity	Coefficients undefined or unstable
A4	Zero Conditional Mean	Biased & inconsistent estimates
A5	Homoskedasticity	Inefficient estimates; invalid SEs
A6	Normality of Errors	Invalid inference in small samples

A1 — Linearity

The relationship between X and Y must be **linear in parameters.**

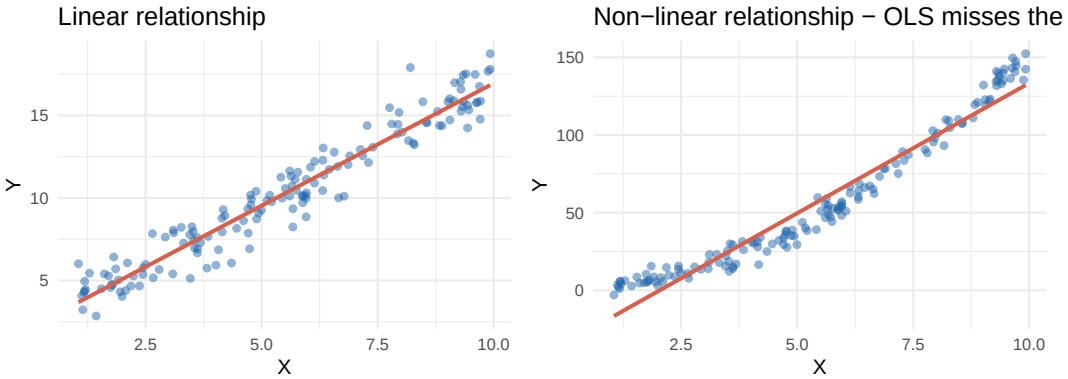


Figure 3

A4 — Zero Conditional Mean (Exogeneity)

$$E[\varepsilon \mid X] = 0$$

The error term must be **uncorrelated with the regressors.**

Why it matters

If an omitted variable affects both X and Y, the coefficient on X absorbs that relationship — giving us a **biased estimate.**

$$\hat{\beta}_{\text{OLS}} \neq \beta_{\text{true}}$$

Common causes

- Omitted variable bias
- Reverse causality
- Measurement error in X

Common fixes

- Add controls
- Instrumental variables
- Fixed effects / DiD

Zero Conditional Mean (Exogeneity)

This is the most consequential assumption. No diagnostic test can confirm it — it requires **theoretical justification.**

In reality, we often don't have exogeneity. Thats ok if we acklowledge not a cuasal relationship

Common Regression Problems

Multicollinearity — What It Is

Definition: Two or more regressors are highly (linearly) correlated.

Why it's a problem

- OLS cannot separately identify each coefficient
- Standard errors **inflate** $\rightarrow$ t-stats shrink
- Coefficients become unstable across samples
- Estimates are still *unbiased*, but *imprecise*

Common triggers in policy research

- GDP & GDP per capita in the same model
- Population & total spending (vs. per capita)
- Highly correlated index components
- Dummy variables that nearly exhaust categories

Perfect multicollinearity (e.g., two identical variables) makes OLS undefined. Ordinary high correlation is *multicollinearity* — more common and fixable.

Detecting Multicollinearity

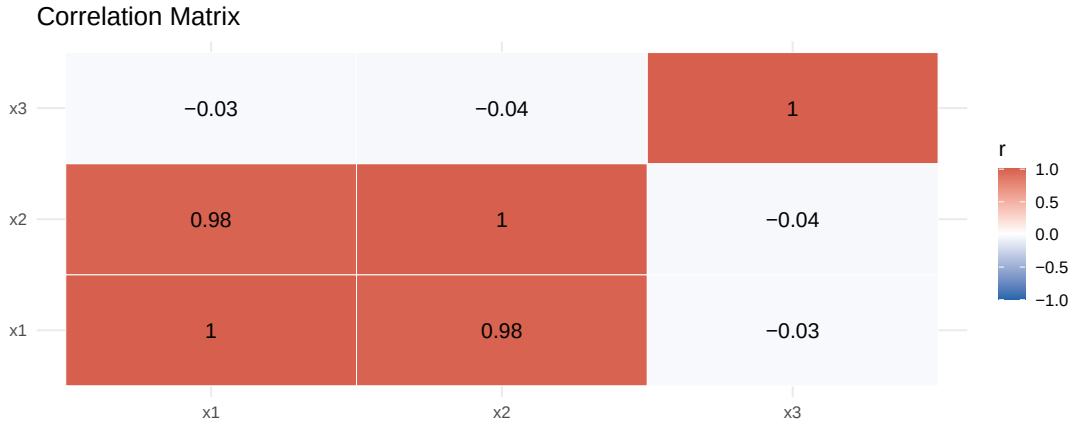


Figure 4

Detecting Multicollinearity — VIF

The **Variance Inflation Factor** quantifies how much a coefficient's variance is inflated by collinearity.

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 from regressing X_j on all other predictors.

Table 5

Variable	VIF	Interpretation
x1	25.06	Severe — action needed
x2	25.08	Severe — action needed
x3	1.00	Acceptable

Addressing Multicollinearity

Strategy	When to use
Drop one of the correlated variables	If variables measure the same concept
Use a composite / index	If both variables belong theoretically
Center or standardize variables	For interaction terms specifically
Ridge / regularised regression	When dropping is theoretically unjustifiable
More data	If sample is small relative to predictors

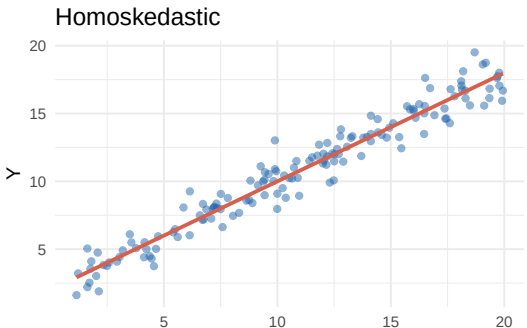
Dropping a variable introduces **omitted variable bias** if it belongs in the model. Multicollinearity is a data problem; OVB is a model problem — don't trade one for the other carelessly.

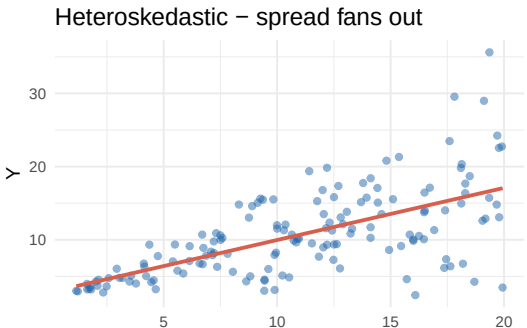
Heteroskedasticity — What It Is

Definition: The variance of the error term is **not constant** across observations.

$$\text{Homoskedastic: } \text{Var}(\varepsilon_i \mid X_i) = \sigma^2 \quad \forall i$$

$$\text{Heteroskedastic: } \text{Var}(\varepsilon_i \mid X_i) = \sigma_i^2$$





Detecting Heteroskedasticity

Visual: Residual vs. Fitted plot — look for a funnel or pattern.

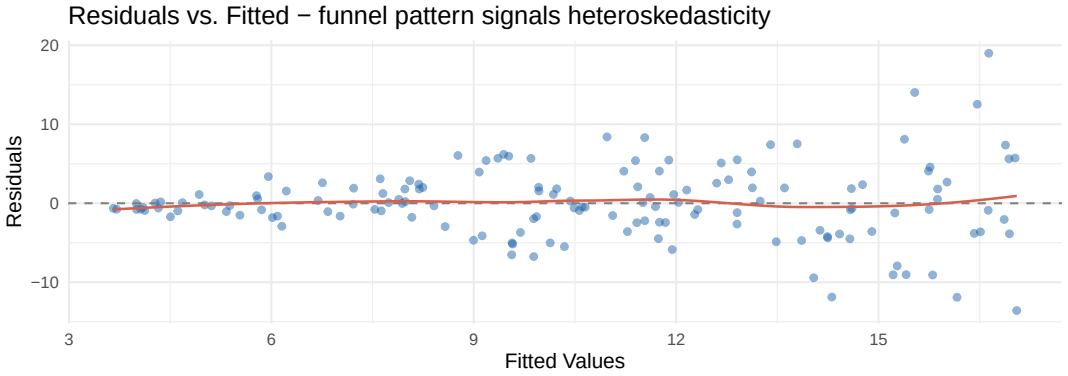


Figure 6

Formal test: Breusch-Pagan test (H_0: homoskedasticity)

Fixing Heteroskedasticity — Robust Standard Errors

OLS coefficients remain **unbiased**, but standard errors are wrong.

Solution: Use heteroskedasticity-consistent (HC) standard errors.

Table 6

	Standard	Robust
	OLS (default SE)	HC3 Robust SE
(Intercept)	2.840***	2.840***
	(0.861)	(0.660)
X	0.713***	0.713***
	(0.071)	(0.082)
Num.Obs.	150	150
R2	0.408	0.408

* p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01

Autocorrelation — What It Is

Definition: Error terms are correlated **across observations**, typically over time.

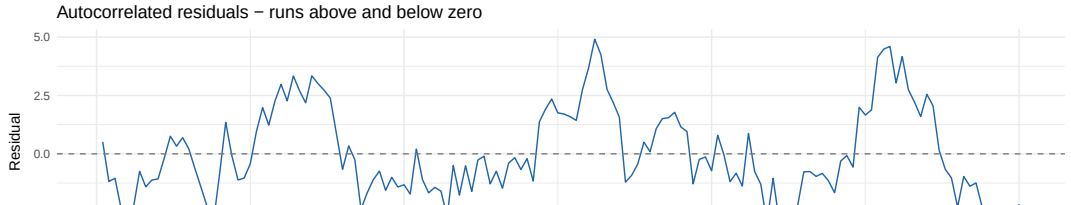
$$\text{Cov}(\varepsilon_t, \varepsilon_{t-k}) \neq 0 \quad \text{for some } k \neq 0$$

Why it's a problem

- Standard errors are **underestimated**
- t-stats are inflated $\rightarrow$ false positives
- OLS is no longer the most efficient estimator

Common in

- Time series (GDP, prices, conflict)
- Panel data (country-years, surveys)
- Spatial data (neighbouring units)



Detecting Autocorrelation

ACF plot — bars outside the blue band indicate significant autocorrelation.

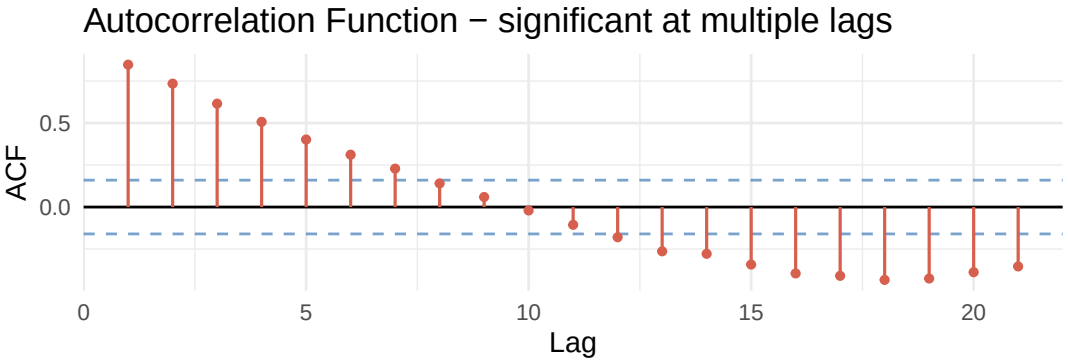


Figure 8

Durbin-Watson test (H_0: no first-order autocorrelation; DW 2 is ideal)

```
dwtest(m_tg)
```

Fixing Autocorrelation

Strategy	When to use
Newey-West (HAC) standard errors	Most common; adjusts SE for AC + hetero
Cluster-robust SE (by unit or time)	Panel data with within-unit correlation
Add lagged DV as a control	Captures temporal persistence in Y
FGLS / Prais-Winsten	When efficiency gain matters more than robustness
First-differencing	Unit-root processes; removes trend-driven AC

Fixing Autocorrelation

Table 7

	Unadjusted	HAC-adjusted
	OLS	Newey-West
(Intercept)	1.130***	1.130***
	(0.315)	(0.291)
Time Trend	0.500***	0.500***
	(0.004)	(0.004)
Num.Obs.	150	150
R2	0.992	0.992

* p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01

Questions

Questions

. . . before we move on to Experiments

Where We Left Off in Lecture 2

- The fundamental problem of causal inference: we never observe both $Y_i(1)$ and $Y_i(0)$ for the same unit
- Causal inference is, at its core, a strategy for approximating the missing counterfactual
- Lecture 2 previewed **random assignment** as the "gold standard" and four workarounds when you can't randomize: DiD, IV, RD, matching
- Today we go *deep* on the gold standard itself — what makes it work, where it comes from, and where it breaks down

Why Experiments? The Logic of Randomized Comparison

The Core Idea

- An **experiment**: the researcher deliberately assigns some units to a **treatment group** and others to a **control group**, then compares outcomes
- "Experiments, or RCTs (randomized control trials) as they are often called, involve the randomized assignment of individuals into one of two groups [...] a treatment group that receives some intervention; and a control group that does not." *(EMPS Ch. 7)*
- Most real experiments have *several* treatment arms, and sometimes more than one control arm

A Surprisingly Recent Idea

- The first recorded experiment: Nuremberg, 1835 — testing a homeopathic remedy (diluted salt in water) on 100 residents split into treatment and control
- Political science's first experiment came nearly a century later: Harold Gosnell's 1924 study mailing GOTV postcards to randomly chosen wards
- Political science did not take experiments seriously as a mainstream method until the 1980s

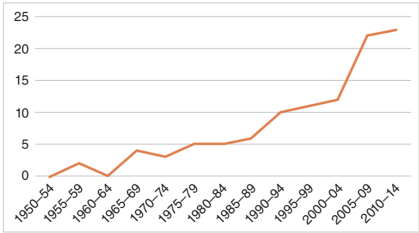


Figure 1 Experimental Articles in APSR, 1950–2014. (Note: Between 1906 and 1954, no experimental articles appeared.)

Why Randomization Works

- Randomization creates **exogeneity**: assignment to treatment becomes statistically unrelated to any confounder, observed or unobserved
- What covariates must we control for in a well-randomized experiment? **None** — treatment and control groups are balanced *in expectation* on every covariate, seen or unseen
- This is the design-based alternative to "controlling for everything" in a regression — the foundation of every identification strategy we discuss the rest of the course

The Experimental Model

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

- D_i: the treatment indicator — $D_i = 1$ if treated, $D_i = 0$ if control
- β_1 is the **Average Treatment Effect (ATE)**: the average difference in outcomes between treated and control groups
- A simple bivariate regression — no controls needed, given proper randomization and adequate sample size

Random Assignment vs. Random Sampling

Two Different Coin Flips

These two ideas get conflated constantly, but they answer *completely different questions*:

Random Sampling

- **Who is in your study at all?**
- Drawing a sample so every member of a population has a known, nonzero chance of selection
- Governs whether your *sample* represents your *population* — this is an **external validity** concern

Random Assignment

- **Once you have your sample, who gets the treatment?**
- Randomly sorting your existing sample into treatment and control arms
- Governs whether your treatment/control comparison is causally clean — this is an **internal validity** concern

Why the Distinction Matters

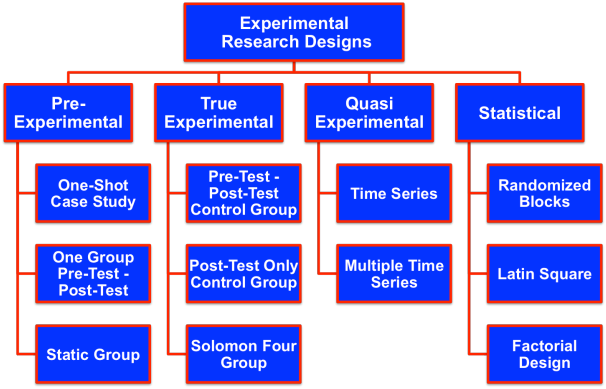
- "It is important to differentiate the random assignment necessary to conduct an experiment from [...] random sampling." *(EMPS Ch. 7)*
- A consequence: how the sample was *collected* matters less than whether assignment to groups *within that sample* was random
- This means a good experiment can run on a **convenience sample** — a sample of easy-to-reach people (newspaper ads, online panels, Mechanical Turk)
- Convenience samples don't undermine internal validity, but they sharply limit how far you can *generalize* the result

Check Your Intuition

- "Is 'random assignment' another way of saying 'random sampling' and vice versa? If not, how are they distinct from one another?" *(EMPS Ch. 7, Check-in Question 1)*
- You could have: a perfectly random national sample with *no* random assignment (an observational survey)
- Or: a convenience sample of college sophomores with *perfect* random assignment to treatment/control (a classic lab experiment)
- Each gets you something different — and a study can be strong on one and weak on the other

Types of Experiments

Four Families, One Logic



- All four types share the same backbone — random assignment to treatment vs. control
- They differ in *where* the randomization happens and *how much control* the researcher has over the setting

Type 1: Survey Experiments

- A **survey experiment**: an experiment embedded inside a survey instrument — participants answer questions as in any survey, but at some point are randomly assigned to a treatment or control condition
- Hugely efficient: no physical space needed, cheap online panels (YouGov, Qualtrics, MTurk), large samples for modest cost
- Limitation: treatments are usually text-based (a paragraph to read), and outcomes are self-reported attitudes — you can't always verify respondents actually absorbed the manipulation

Worked Example: Framing Trade Policy

Sensitivity to Issue Framing on Trade Policy Preferences: Evidence from a Survey Experiment

Martin Ardanaz, M. Victoria Murillo, and Pablo M. Pinto

- Ardanaz, Murillo & Pinto fielded a survey experiment in Argentina on attitudes toward trade
- Respondents were randomly assigned to one of four framing conditions before being asked about support for increasing trade:

- Group 1 (25 percent)—pro-trade introduction: "Some people believe that increasing trade with other nations creates jobs and allows you to buy goods and services at lower prices."
- Group 2 (25 percent)—anti-trade introduction: "Some people believe that increasing trade with other nations causes unemployment and hurts Argentine producers."
- Group 3 (25 percent)—both introductions.

Worked Example: Results

TABLE 2. *Percentages of respondents who favor increasing trade*

All respondents (N = 2,729)	71.31
Pro-trade introduction (N = 687)	69.43
Anti-trade introduction (N = 674)	67.95
Both introductions (N = 687)	69.29
No introduction (Group 4: N = 673)	78.46

Note: Table shows percentage who strongly agree and somewhat agree with the question.

- Support for increased trade varied substantially by framing — highest with *no* introduction at all (78.5%), lowest after the anti-trade frame (68.0%)
- Note how cheaply this was fielded: $N = 2{,}729$ respondents across four arms, no lab space, no field staff

Type 2: Laboratory Experiments

- A **laboratory experiment** takes place in a controlled environment the researcher manages directly — near-complete control over delivery and measurement
- Big advantages over survey experiments: researchers can verify real engagement, observe actual behavior (not just self-report), and build richer, more realistic interactions
- Big cost: participants must physically travel to the lab — samples are usually small (100–200), shrinking statistical power

Worked Example: Partisanship Under Pressure

- White, Laird & Allen (2014) ran a lab experiment at a historically Black college two months before the 2012 Obama–Romney election
- Participants privately allocated a hypothetical \$100 between the two candidates
- Some were randomly told that for every \$10 given to Romney, *they personally* would receive \$1 in cash — a direct economic incentive to defect from their party
- A second treatment arm added: donation amounts would be made public in the campus newspaper

Worked Example: What They Found

- **Control group**: averaged \$90 of \$100 to Obama
- **Private economic incentive**: average dropped to \$68 to Obama
- **Incentive + public exposure**: average rose back to \$86 to Obama

- The threat of group exposure pulled behavior back toward the partisan norm — a social-pressure effect that would be very hard to engineer convincingly in a survey
- This kind of manipulated, observable social interaction is exactly what a lab buys you over a survey instrument

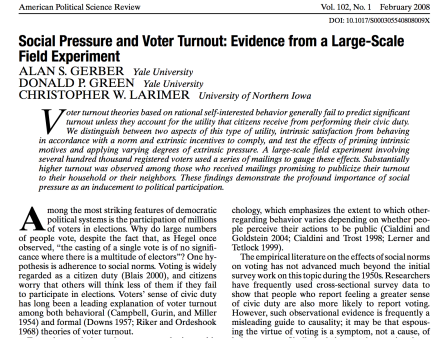
Type 3: Field Experiments

- A **field experiment**: randomization is still researcher-controlled, but the intervention and outcome occur in subjects' real-world environment, not a lab or survey screen
- Strongest combination of internal *and* external validity among the four types — you are watching real behavior, not a stated preference
- Costly: often requires field staff, printed materials, mailings, or partner organizations, and the timeline is dictated by real-world events (you can't reschedule an election)

Worked Example: Mobilization and Knowledge

- Shineman (2018) ran a field experiment around a 2011 San Francisco municipal election
- 178 subjects completed a pre-election survey, were randomly assigned to receive different GOTV mobilization treatments (including voter-registration assistance) or nothing, then completed a post-election survey
- Finding: mobilization didn't just raise turnout — it also raised measured political *knowledge* after the election
- That second finding was only possible because the design followed real subjects through a real election, not a hypothetical one

A Related Field Classic: Social Pressure and Turnout



American Political Science Review Vol. 102, No. 1 February 2008
DOI: 10.1017/S000305540808009X

Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment

ALAN S. GERBER *Yale University*
DONALD P. GREEN *Yale University*
CHRISTOPHER W. LARIMER *University of Northern Iowa*

Voter turnout theories based on rational self-interested behavior generally fail to predict significant turnout unless they account for the utility that citizens receive from performing their civic duty. We distinguish between two aspects of this type of utility, intrinsic satisfaction from behaving in accordance with a norm and extrinsic incentives to comply, and test the effects of priming intrinsic motives and applying varying degrees of extrinsic pressure. A large-scale field experiment involving several hundred thousand registered voters used a series of mailings to gauge these effects. Substantially higher turnout was observed among those who received mailings promising to publicize their turnout to their household or their neighbors. These findings demonstrate the profound importance of social pressure as an inducement to political participation.

Among the most striking features of democratic political systems is the participation of millions of voters in elections. Why do large numbers of people vote, despite the fact that, as Hegel once observed, "the casting of a single vote is of no significance where there is a multitude of electors"? One hypothesis is adherence to social norms. Voting is widely regarded as a citizen duty (Blais 2000), and citizens worry that others will think less of them if they fail to participate in elections. Voters' sense of civic duty has long been a leading explanation of voter turnout among both behavioral (Campbell, Gurin, and Miller 1954) and formal (Downs 1957; Riker and Ordeshook 1968) theories of voter turnout.

chology, which emphasizes the extent to which other-regarding behavior varies depending on whether people perceive their actions to be public (Cialdini and Goldstein 2004; Cialdini and Trost 1998; Lerner and Tetlock 1999).

The empirical literature on the effects of social norms on voting has not advanced much beyond the initial survey work on this topic during the 1950s. Researchers have frequently used cross-sectional survey data to show that people who report feeling a greater sense of civic duty are also more likely to report voting. However, such observational evidence is frequently a misleading guide to causality; it may be that espousing the virtue of voting is a symptom, not a cause, of

- Gerber, Green & Larimer (2008) randomly assigned 344,000+ Michigan households to receive different GOTV mailers before a primary — including one showing neighbors each other's past voting records
- A massive field experiment built directly on a partner relationship with the state's voter file

Checking the Randomization Worked

TABLE 1. Relationship between Treatment Group Assignment and Covariates (Household-Level Data)

	Control	Civic Duty	Hawthorne	Self	Neighbors
	Mean	Mean	Mean	Mean	Mean
Household size	1.91	1.91	1.91	1.91	1.91
Nov 2002	.83	.84	.84	.84	.84
Nov 2000	.87	.87	.87	.86	.87
Aug 2004	.42	.42	.42	.42	.42
Aug 2002	.41	.41	.41	.41	.41
Aug 2000	.26	.27	.26	.26	.26
Female	.50	.50	.50	.50	.50
Age (in years)	51.98	51.85	51.87	51.91	52.01
N =	99,999	20,001	20,002	20,000	20,000

Note: Only registered voters who voted in November 2004 were selected for our sample. Although not included in the table, there were no significant differences between treatment group assignment and covariates measuring race and ethnicity.

- Before trusting any treatment effect, check balance: does random assignment actually produce comparable groups on observables?
- Here, household size, past turnout, gender, and age are nearly identical across all five arms — exactly what good randomization should produce

The Payoff: Estimated Effects

TABLE 3. OLS Regression Estimates of the Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Model Specifications		
	(a)	(b)	(c)
Civic Duty Treatment (Robust cluster standard errors)	.018* (.003)	.018* (.003)	.018* (.003)
Hawthorne Treatment (Robust cluster standard errors)	.026* (.003)	.026* (.003)	.025* (.003)
Self-Treatment (Robust cluster standard errors)	.049* (.003)	.049* (.003)	.048* (.003)
Neighbors Treatment (Robust cluster standard errors)	.081* (.003)	.082* (.003)	.081* (.003)
N of individuals	344,084	344,084	344,084
Covariates**	No	No	Yes
Block-level fixed effects	No	Yes	Yes

Note: Blocks refer to clusters of neighboring voters within which random assignment occurred. Robust cluster standard errors account for the clustering of individuals within household, which was the unit of random assignment.
* $p < .001$.
** Covariates are dummy variables for voting in general elections in November 2002 and 2000, primary elections in August 2004, 2002, and 2000.

- The "Neighbors" mailer — the one disclosing household voting records to neighbors — raised turnout by roughly 8 percentage points over control
- Estimates are stable across model specifications (a), (b), (c) — a sign the effect isn't an artifact of one particular regression choice

Type 4: Natural Experiments

- A **natural experiment**: the randomization wasn't designed by the researcher at all — some outside process (an administrative rule, a redrawn border, an "act of God") created **as-if random** variation in treatment
- "No matter how the randomization happened, if it was not the researcher who did it, then it is a natural experiment." *(EMPS Ch. 7)*
- The payoff when one is found: realism as high as the real world itself, since you're observing actual people facing an actual policy
- The catch: you must go looking for one, and you can never be fully certain the "randomization" was truly as-if random

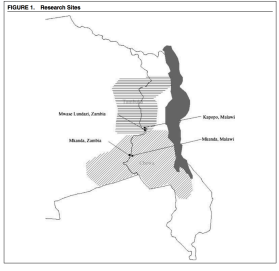
Natural Experiments, Closer to Home: Posner (2004)

The Political Salience of Cultural Difference: Why Chewas and Tumbukas Are Allies in Zambia and Adversaries in Malawi

DANIEL N. POSNER *University of California, Los Angeles*

This paper explores the conditions under which cultural cleavages become politically salient. It does so by taking advantage of the natural experiment afforded by the division of the Chewa and Tumbuka peoples by the border between Zambia and Malawi. I document that, while the objective cultural differences between Chewas and Tumbukas on both sides of the border are identical, the political salience of the division between these communities is altogether different. I argue that this difference stems from the different sizes of the Chewa and Tumbuka communities in each country relative to each country's national political arena. In Malawi, Chewas and Tumbukas are each large groups vis-à-vis the country as a whole and, thus, serve as viable bases for political coalition-building. In Zambia, Chewas and Tumbukas are small relative to the country as a whole and, thus, not useful to mobilize as bases of political support. The analysis suggests that the political salience of a cultural cleavage depends not on the nature of the cleavage itself (since it is identical in both countries) but on the sizes of the groups it defines and whether or not they will be useful vehicles for political competition.

FIGURE 1. Research Sites



A Border as a Treatment

- Posner compares two ethnic groups — Chewas and Tumbukas — split across the Zambia–Malawi border by colonial-era boundary drawing
- The same two ethnic groups behave very differently politically depending on which side of the border they ended up on
- The border itself functioned like an as-if random assignment of *national institutional context*, while holding the ethnic groups constant
- This is the natural-experiment logic at its cleanest: find a discontinuity nobody designed, and let it do the randomizing for you

Internal, External, and Construct Validity

Three Validities, One Trade-Off

EMPS frames these as a three-way balancing act every experimenter has to navigate — you rarely maximize all three at once:

- **Internal validity**: did the treatment actually *cause* the outcome we observe, within this study?
- **External validity**: would the finding generalize beyond this sample, setting, and moment?
- **Construct validity**: does the manipulation in the experiment actually capture the theoretical concept you care about?

Construct Validity: The One We Skip Too Often

- "A great deal of work must be taken on the front-end to ensure good construct validity, or the ability of the experiment to actually speak to the theory or research question at hand." *(EMPS Ch. 7)*
- Easy to design a clever, internally valid experiment that doesn't actually test your theory
- Maimonides' Rule again: a clean cutoff design that, on inspection, was arguably measuring *parental resourcefulness*, not class size
- Construct validity is a design question, not a statistical one — no significance test can rescue a manipulation that doesn't match the concept

How the Four Types Stack Up

	Lab	Field	Survey	Natural
Construct validity	depends	depends	depends	depends
Internal validity	high	depends	high	low*
External validity	low	high	high	high*

*(EMPS Ch. 7, Table 1; = "it depends," case by case, for natural experiments)**

- Construct validity is "depends" across the board — it lives in design quality, not experiment type
- Internal and external validity trade off in a fairly predictable pattern across the other three types

A Useful Mnemonic: Trading Control for Realism

More Control

- Lab experiments
- Survey experiments
- Higher internal validity
- Often lower external validity

More Realism

- Field experiments
- Natural experiments
- Often higher external validity
- Internal validity depends on design quality

- Recent evidence pushes back a bit on this stereotype: some survey-experiment findings have been shown to approximate real-world behavior reasonably well *(Hainmueller, Hangartner & Yamamoto 2015, cited in EMPS Ch. 7)*

Advantages and Disadvantages of Experiments

Why Experiments Win on Causality

- "No other research method in this textbook is quite as good as a simple experiment at assessing causality or at achieving good internal validity." *(EMPS Ch. 7)*
- The statistics required are often minimal — a comparison of group averages, not a complicated model
- Versatile: as long as you can randomize participants into groups and measure an outcome, you can run *some* version of an experiment
- This is exactly why Lecture 2 called randomization the identification strategy every other strategy (DiD, IV, RD, matching) is trying to *approximate*

Why Experiments Aren't a Free Lunch

- **Weak external validity, often**: artificial settings (a survey scale, a lab room) may not reflect how people actually behave or think
- **Feasibility and cost**: field and lab experiments require money, staff, time, and sometimes a partner organization
- **Ethics**: you can't always *ethically* randomize the thing you care about (withholding a beneficial policy from a control group, for instance)
- **Snapshots, not repeatable measurements**: most experiments run once, on one sample, in one political moment — the same experiment is rarely run again on the same population

When an Experiment Isn't the Right Tool

- Many of the policy questions you'll care about most can't be ethically or feasibly randomized: democratization, war, constitutional design, long-run economic development
- This is *exactly* the gap that Lecture 2's identification strategies are built to fill:
  - **Difference-in-Differences**, **Instrumental Variables**, **Regression Discontinuity**, **Matching**
- Natural experiments sit at the boundary — they borrow the logic of randomization from a process the researcher didn't design
- *Lack of time or money*

Threats to a Clean Experiment

Attrition

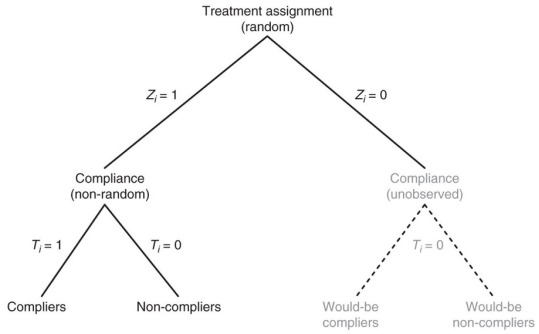
- Units **drop out** before we observe their outcome
- A problem when attrition is *not random* — e.g., frustrated or busy participants leave disproportionately from one arm
- This undermines the exogeneity that randomization was supposed to buy you

Balance

- Do treatment and control groups look the same on other (observed) variables, on average?
- Checked with a **balance table**: comparing means and variances of covariates across arms (recall the Gerber, Green & Larimer table from a few slides ago)
- **Blocking**: pre-assigning units to ensure balance on key covariates — especially valuable with small samples

Compliance

- Do units actually receive the treatment they were assigned to?

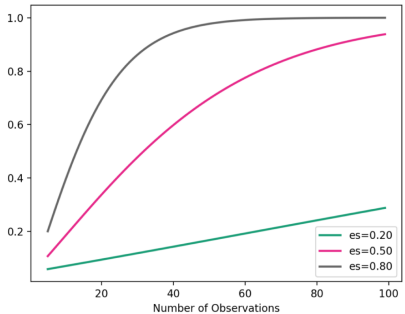


- **Never-takers**, **always-takers**, **compliers**, and (rarely) **defiers**
- Solution: an **intent-to-treat (ITT)** analysis — compare groups *by assignment*, regardless of whether they actually took the treatment. Non-compliance shrinks the estimated ATE toward zero, like added noise

Power, Pre-Registration & Research Integrity

Power Analysis

- **Statistical power**: the probability of detecting a true effect, if one exists — correctly rejecting a false null
- Power depends on sample size, expected effect size, and your chosen significance level (α)



- Run a power analysis *before* collecting data — an underpowered experiment can't

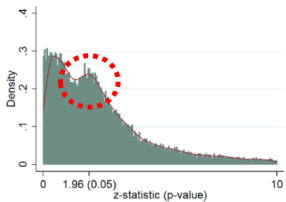
The Replication Crisis and p-Hacking

- Exploratory "fishing" for significant results inflates false positives across a field
- p-hacking: trying many specifications or subgroups until something crosses $p < 0.05$

Economics

Brodeur et al (AEJ:A, in press)

"Star Wars: The empirics strike back"

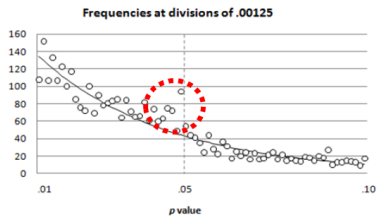


(b) De-rounded distribution of z-statistics.

Psychology

Masicampo Lalande (QJEP, 2012)

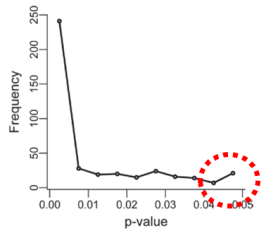
"A peculiar prevalence of p values just below .05"



Biology

Head et al (PLOS Biology 2015)

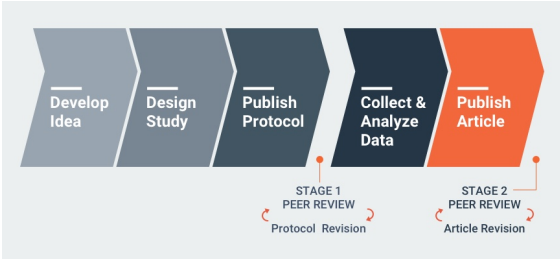
"Extent and Consequences of P-Hacking in Science"



- Suspicious clustering of p-values just under 0.05, across economics, psychology, and biology — a signature of selective reporting, not nature

Pre-Registration

- A **pre-analysis plan**: specifying hypotheses, sample, and analysis *before* seeing the outcome data
- Prevents exploratory hypothesis-shopping from being presented as confirmatory testing



- **Registered reports**: journals review and provisionally accept a study based on its design alone, before results exist
- Resources: EGAP (egap.org), the Center for Open Science (cos.io)

Closing the Loop: Experiments and Identification

Back to Lecture 2's Toolkit

- Lecture 2 introduced random assignment as the gold standard, then four workarounds for when you can't randomize: DiD, IV, RD, matching
- Today's deeper dive should make clear *why* those workarounds exist: each one tries to recreate some piece of what random assignment gives you for free
- An **instrumental variable** is, in effect, a search for a natural source of as-if random variation — the same logic behind every natural experiment we discussed today
- **Regression discontinuity** formalizes exactly the Maimonides'-Rule-style cutoff logic — compare units just above and just below an arbitrary threshold

A Unified Way to See the Course So Far

- Every method in this course is answering the same question: *can I trust that my comparison group approximates the missing counterfactual?*
- Experiments answer "yes" by design, when randomization is done well and holds up under attrition, balance, and compliance checks
- Quasi-experimental strategies answer "probably," under explicit, checkable assumptions
- Observational comparisons without either need the most caution — and the most theoretical justification for why X and ϵ might still be unrelated

Key Terms from Today

- Treatment / control group
- Random assignment vs. random sampling
- Average Treatment Effect (ATE)
- Survey / lab / field / natural experiment
- Convenience sample
- Internal / external / construct validity
- Attrition / balance / compliance
- Intent-to-treat (ITT)
- Pre-registration / power analysis
- Spurious relationship

Before Next Session

Lecture 10: Qualitative Methods 1

- Read: QRM, Ch. 1–3 (foundational chapters on qualitative inquiry)
- Read: EMPS, Ch. 9 (*Small-N and Comparative Designs*)
- Bring a question: what could qualitative methods add to a topic you'd normally study quantitatively?

Questions?

Thank you — see you in the Practicum this afternoon.

See You in the Practicum