(Week 5) Causal inference and the credibility revolution
(Week 6) Quant. analysis: strengths & pitfalls

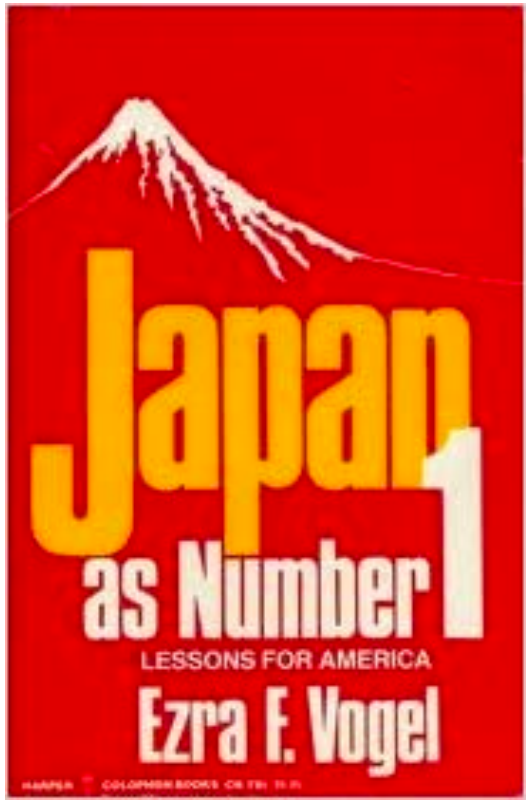
Research design

11 & 18 November, 2019

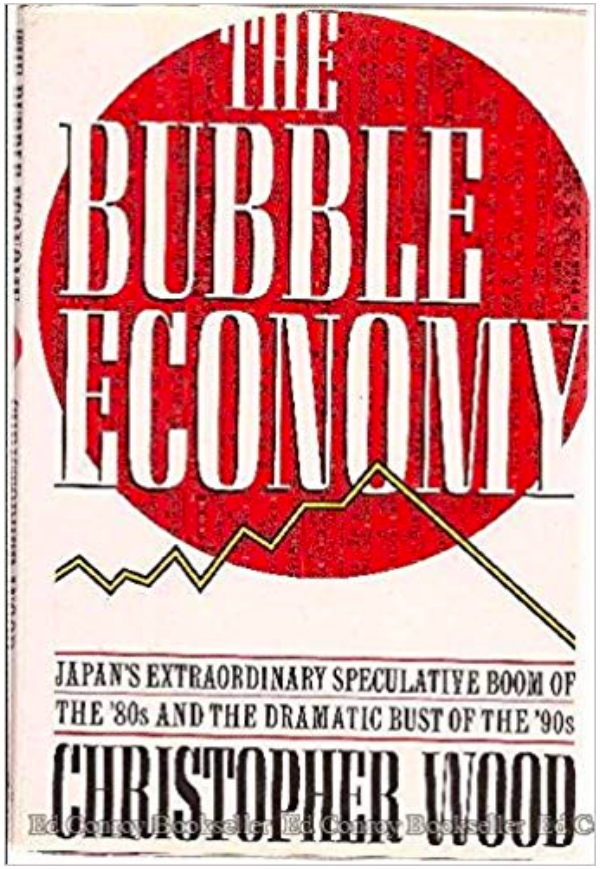
Andrew Eggers

Slides available at http://andy.egge.rs/teaching.html

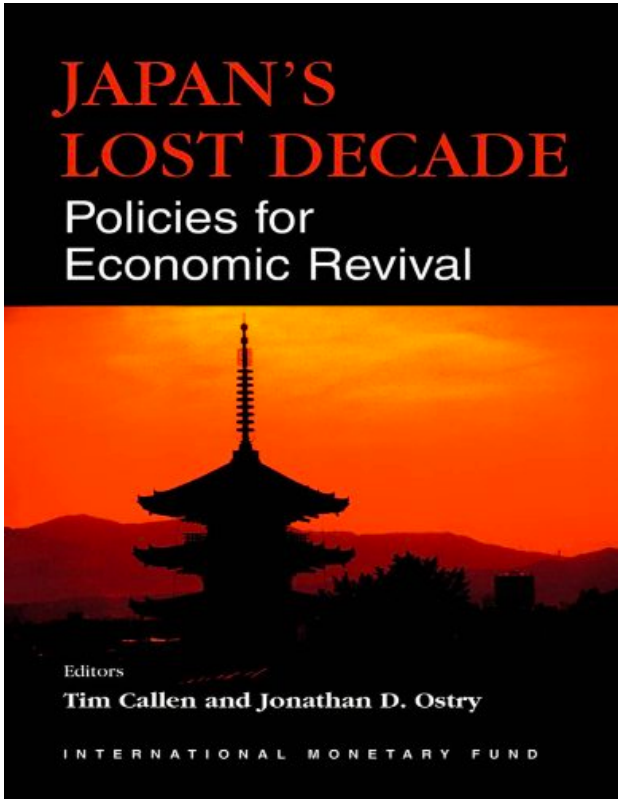
Some personal history



1979



1992



2003

What I work on (recently)

Strategic voting

- How strategically do voters vote?
- Are some types of voters more strategic than others?
- How does strategic voting work in different systems?

Elections and representation

- How does PR affect turnout relative to plurality?
- How does partisanship affect accountability?
- What is the optimal design of a representative legislature?

Money in politics

- Does it matter that members of the US Congress are wealthy?

Methods

- How do we test the assumptions behind our research designs?

Outline of large-N quantitative work

Develop a research question (motivated by puzzle, or ignorance, or policy question) *e.g. "Does proportional representation (PR) increase turnout compared to plurality?"*

Develop a research design: a way of answering this question with data

- **Observational study:** Find cases that vary in the explanatory variable *e.g. elections in several PR and plurality countries*
- **Quasi-experiment or natural experiment:** Find cases that vary pseudo-randomly in the explanatory variable *e.g. turnout in French villages near population cutoff*
- **Experiment:** Create cases that vary randomly in the explanatory variable *e.g. voting experiment in lab*

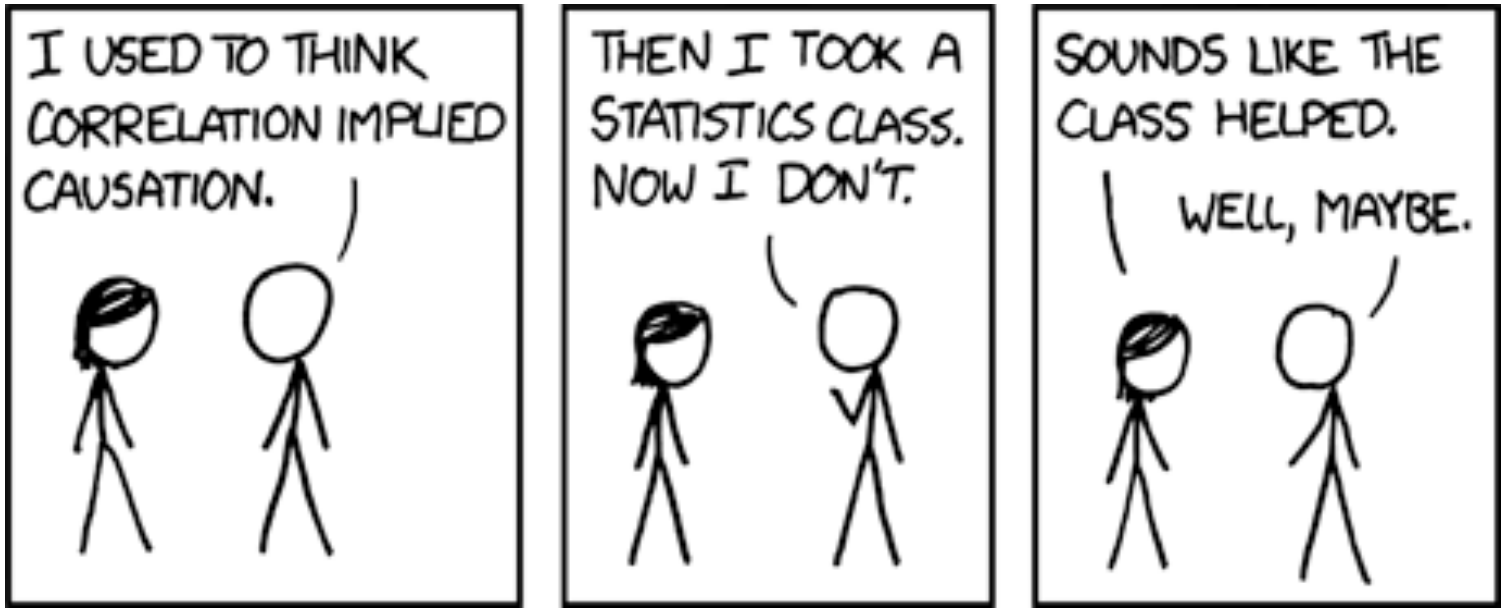
Perform analysis: collect data, run regressions & hypothesis tests

What is large-N quantitative work for?

A regression can be used

- To describe relationships among measures/variables
- To predict outcomes (e.g. forecasting)
- To provide evidence about what caused Y (the outcome)
- To measure the effect of X on Y

Correlation and causation



Correlation doesn't imply causation… except when it does.

Credibility revolution

Increasing awareness across many social science fields that observational studies can provide seriously misleading answers to questions about causality.

A story about program evaluation

National Supported Work Demonstration (1975-1979): ex-offenders, drug addicts, etc. receive 12-18 months of subsidized employment in 10 US cities.

MDRC implementing NSW in 1970s

Does it work? Of 6,600 eligible participants, some randomly assigned to **control group** (no subsidized employment).

	Treatment	Control
Avg earnings after program:	$4,670	$3,819
Avg treatment effect (difference):	$4,670 - $3,819 = $851	

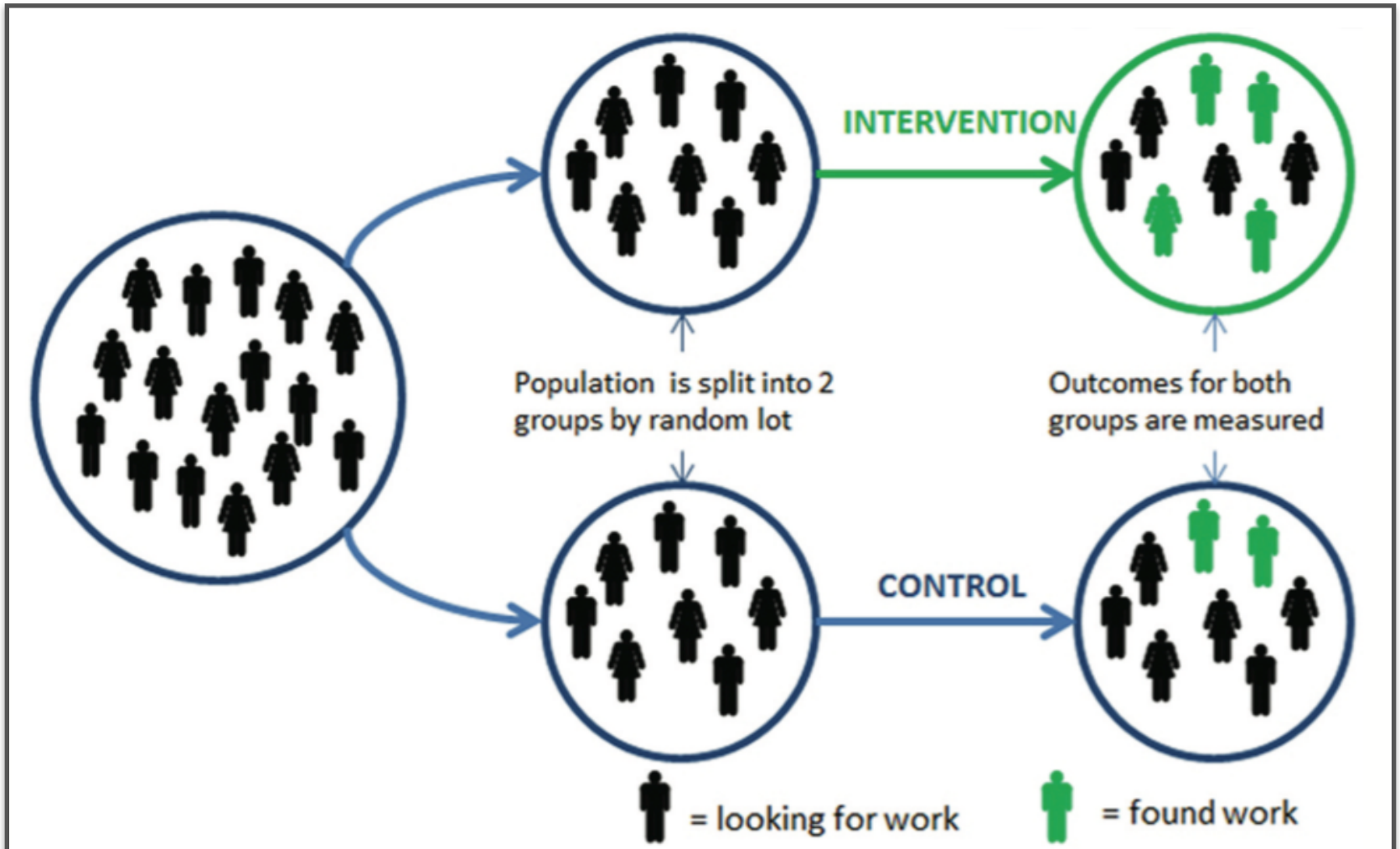


Figure 1. The basic design of a randomised controlled trial (RCT), illustrated with a test of a new 'back to work' programme.

Lalonde (1986): "Evaluating the econometric evaluations of training programs with experimental data"

Robert Lalonde,
University of Chicago

Question: Suppose NSWD had **not** run an experiment. Would typical methods for evaluating observational studies yield the right answer ($851)?

Method: Remove control group from dataset, replace with respondents from typical economic survey. Then run regression like

$$\text{Earnings} = \beta_0 + \beta_1 \text{Training} + \beta_2 \text{Age} + \beta_3 \text{YearsOfSchooling} + \ldots$$

Will β_1 be similar to $\$851$?

Lalonde (1986): "Evaluating the econometric evaluations of training programs with experimental data"

Robert Lalonde,
University of Chicago

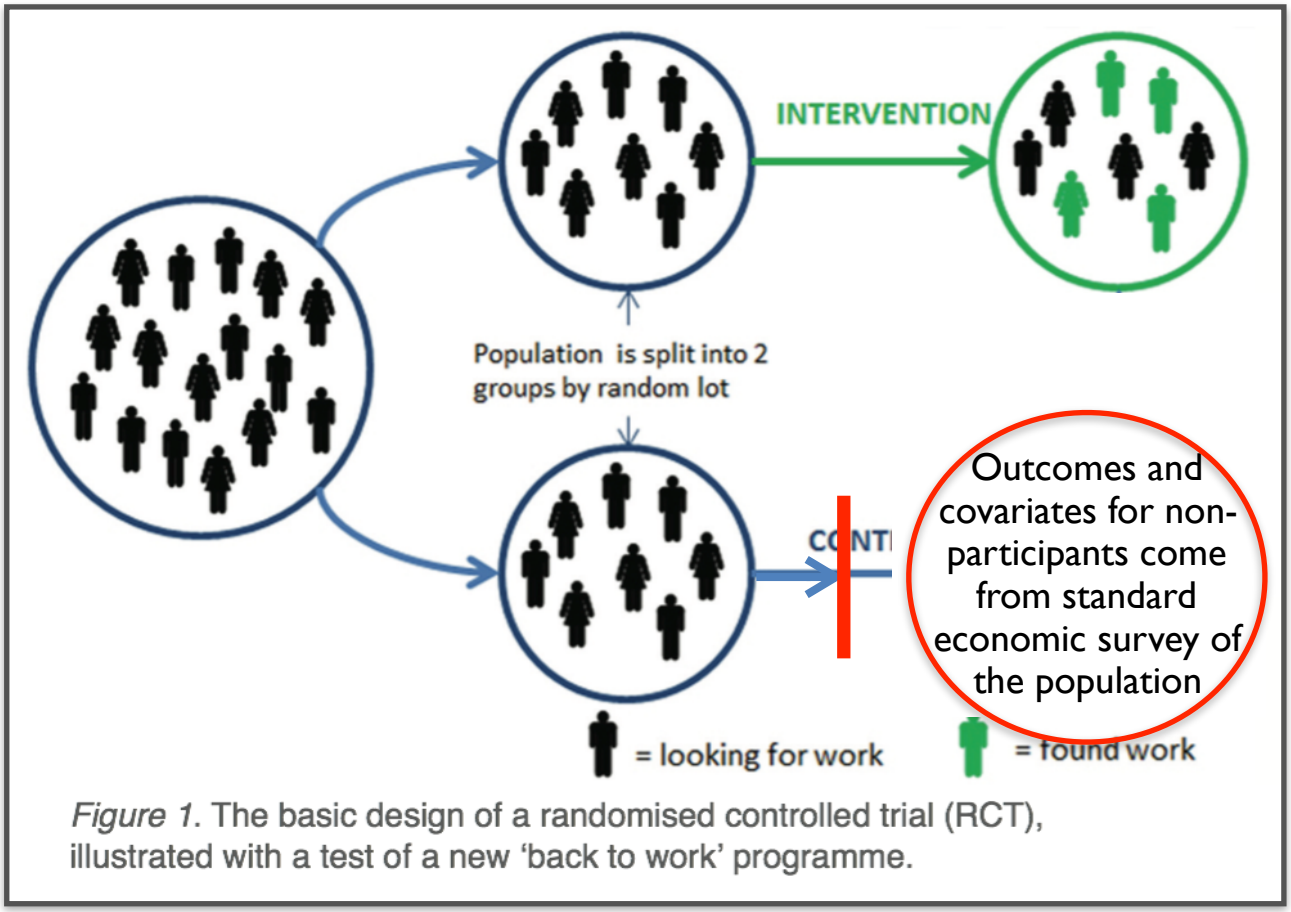


Figure 1. The basic design of a randomised controlled trial (RCT), illustrated with a test of a new 'back to work' programme.

Lalonde (1986): "Evaluating the econometric evaluations of training programs with experimental data"

How close to the experimental benchmark do we get by applying standard econometric approaches to non-experimental data? Not very close!

"Policymakers should be aware that the available non-experimental evaluations of employment and training programs may contain large and unknown biases resulting from specification errors." (p. 617)

But see also Heckman and Hotz (1989), Dehejia and Wahba (1999), Smith and Todd (2001), . . .

What's the problem?

Participants and non-participants differ in important ways that we can't measure or don't know how to control for.

For example?

Is this just a problem with job training programs?

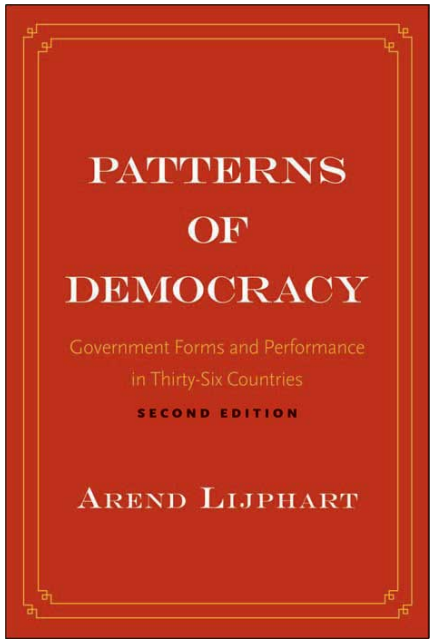
Is this only a problem with large-n quantitative studies?

Fundamental problem of causal inference: we cannot observe the same unit both with and without [job training, WTO membership, democracy], so we have to compare units that may differ in important ways.

Credibility revolution (?)

Critiques like Lalonde's have had a major impact on social sciences — not just evaluation of job training programs.

Before: big questions, cross-country regressions with observational data



How does government form affect performance?

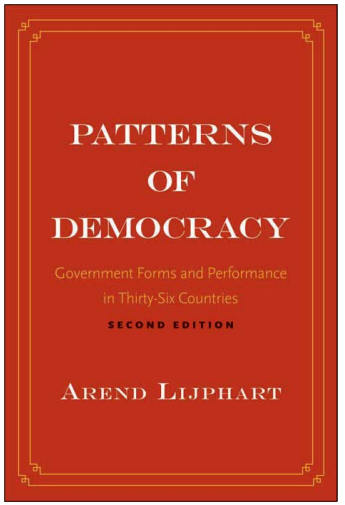
After: narrow questions, experiments and quasi-experiments, "exploiting variation" within one country

(Many talks in Nuffield Politics seminar, IR colloquium)

The rise of the "identification strategy"

Suppose you want to know about the effect of X on Y in population P.

Before: You measure X and Y and some controls Z in (a sample from) population P and run some (possibly complicated) analysis.



How does government form affect performance?

After:

- You run an experiment in which you vary X (or something like X) in a sample that may not look like P; measure difference in means
- You find a sample in which X varies for (conditionally) arbitrary reasons; measure difference in means

Your "identification strategy" is what makes your design more credible than a cross-sectional regression in a sample from P.

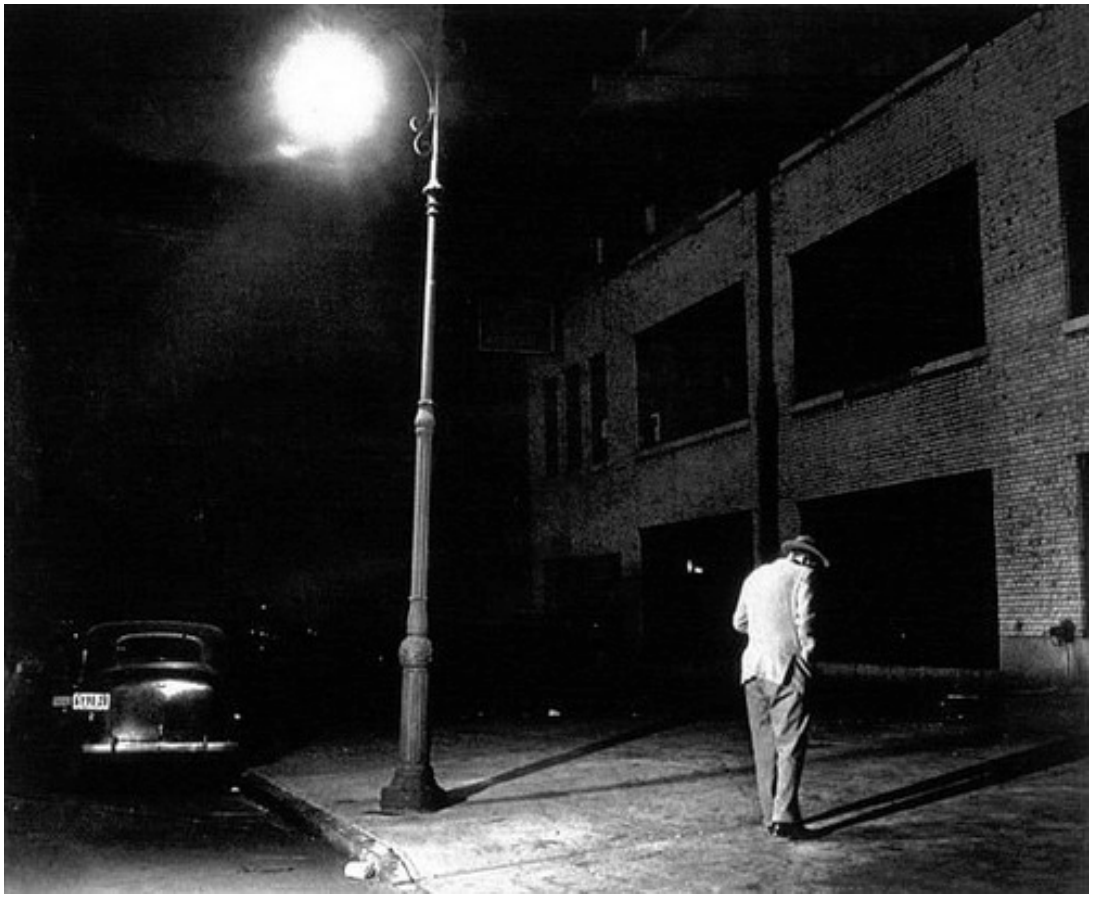
Design vs statistical control

Key feature of design-based approach: choosing or creating settings where **statistical control** is **less necessary**.

Research question	Statistical control approach	(Non-experimental) design approach
What is effect of job training program?	Gather data on a bunch of people including participants and non-participants. Regress wages on participation indicator and controls.	Locate job program that was over-subscribed; compare outcomes for successful and unsuccessful applicants.
What is effect of PR (compared to plurality) on turnout?	Gather data on turnout from various countries. Regress turnout on electoral system indicator and controls.	Compare French cities just above and below population cutoff that determines electoral system.

Looking under a lamppost

Internal validity and external validity — see Cyrus Samii "Causal Empiricism"



What does this mean for you?

Let's back up: what are we trying to do anyway?

When we do research, we are trying to address a (conceptual) **problem**: we are trying to address confusion, ignorance, disagreement about something important. (Not necessarily a **puzzle**!)

The **problem** usually appears in the introduction of a paper, near a word like "but":

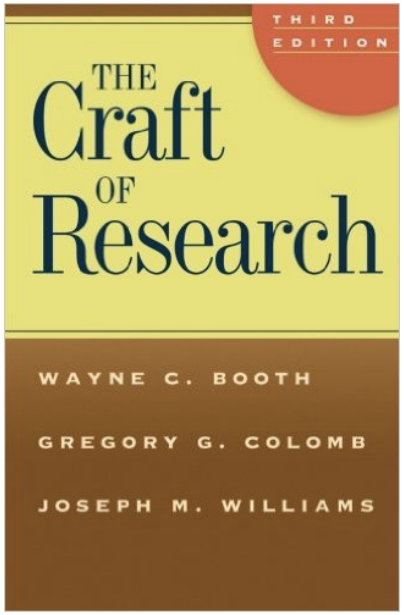
Both theories rest on the assumption that the negative correlation between education levels and anti-immigration sentiment is, at least partly, causal. However, the validity of this underlying premise has not been definitively established. (Cavaillé and Marshall, 2018, APSR)

Comparative political economists generally agree that social democratic parties are the defenders of labor. The persistence of widespread unemployment witnessed under social democratic governments since the early 1970s, **however**, powerfully conflicts with this assumption. (Rueda 2005, APSR)

Also known as "motive": why are you making me read this?

Some types of conceptual problems

- Data does not seem to fit with theory or conventional wisdom about relationship between X and Y
- There are conflicting explanations/ theories for Y
- We don't know how X affects Y ("program evaluation")
- We don't know how Y varies with X (descriptive)



Notes:

- You can start with a **topic** (e.g. "the Left in France", "globalization and preferences"), but eventually you need a **problem**.
- Identifying a problem requires reading the literature. (Gary King: "Whose mind does this change about what?")
- You also need to convince us that it is **important** to address this problem.

So what does this have to do with causal inference and the credibility revolution?

The ingredients of good empirical research:

- a problem (real and consequential)
- a solution: a piece of analysis that addresses the problem

What has changed? The way people assess proposed solutions that rely on empirical measurements of effects.

One response: start by looking for research designs that *randomnistas* will find credible, then see if it addresses a problem.

- Is ballot order randomized in California?
- Does the municipal electoral system depend on an arbitrary population cutoff?
- Is there detailed weather data?

Analyze first, ask questions (i.e. decide on a problem) **later?**

This actually works: there are more interesting questions than credible research designs, so why not start with the design?

And it can work for other types of research too:

- obtain some new data, document some patterns…
- adapt a formal model to a different setting, study its features…

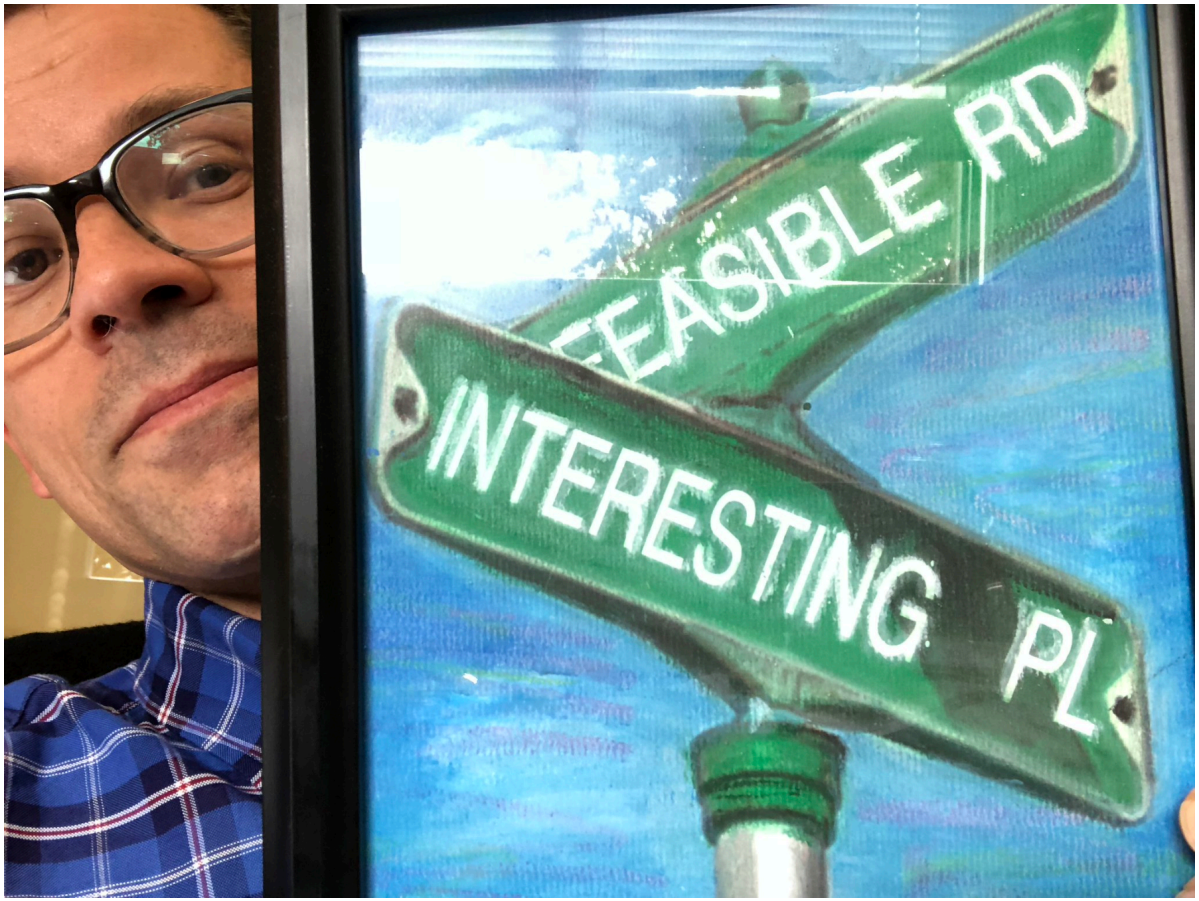
find out if there is something interesting that addresses confusion/ignorance/disagreement in literature.

But the method- or data-driven approach often leads to trivial questions, scattered research profile.

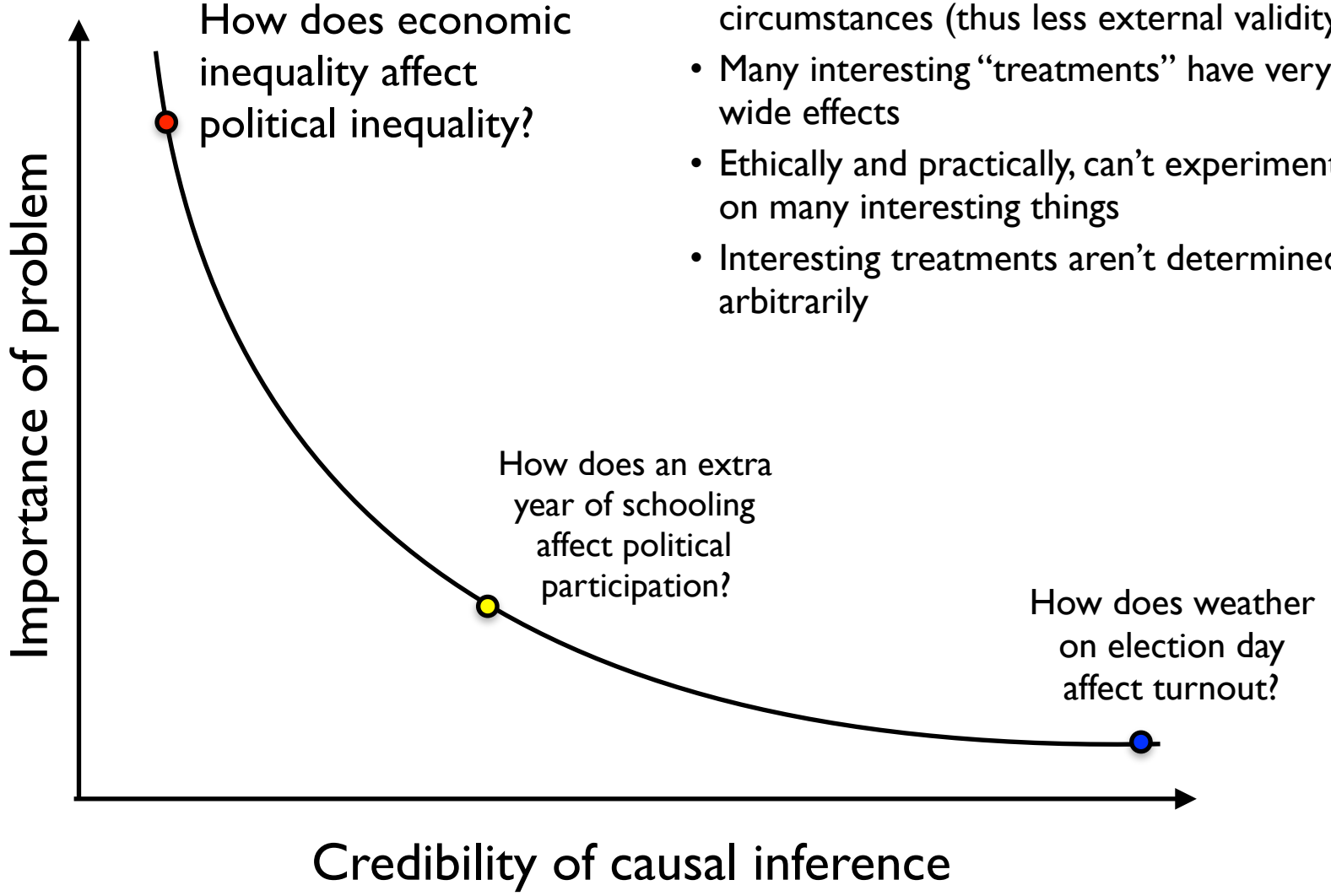
And causal inference-driven approach excludes other types of questions: descriptive, explanatory, conceptual/theoretical.

"Good research occurs at the intersection of interesting and feasible."

Jake Vigdor (U Wash econ)



The credibility/importance tradeoff



Some underlying reasons:

- Internal validity requires special circumstances (thus less external validity)
- Many interesting "treatments" have very wide effects
- Ethically and practically, can't experiment on many interesting things
- Interesting treatments aren't determined arbitrarily

Hypothesis testing

Null hypothesis significance testing (NHST) in social science

Typical **null hypothesis** is "nothing is going on", e.g. two groups have same mean, two measures not related, etc.

Typical (frequentist) procedure:

(1) Calculate test statistic (e.g. difference in means) in reality (i.e. in data).

(2) Estimate/simulate distribution of test statistic if null hypothesis were true (**null distribution**).

(3) Reject null hypothesis if (2) suggests that a result "as adverse to the null hypothesis" as (1) is sufficient unlikely (e.g. probability < .05)

In praise of NHST

- Logically clear
- Even when "repeated sampling" far-fetched, gives standard for saying "coefficient is large compared to uncertainty in model"
- It is standard so you must understand it.

Some problems with NHST: arbitrariness of .05 cutoff

Surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?

Rosnow and Rosenthal, 1989, p. 1277

Some problems with NHST: p-hacking

Because results are "significant" if $p<.05$, researchers try to achieve these results through various methods.

But then does hypothesis testing make any sense?

Four kinds of "search" to worry about

Specification search: Having chosen an X and Y of interest and a setting, try various control variables, functional forms, etc until you find a significant relationship between X and Y

Outcome search: Having found a setting where X is quasi-randomly assigned, try various outcome variables Y until find a significant relationship

Treatment search: Having chosen a Y of interest and a setting, run a regression and choose your hypothesis based on what coefficients turn out to be significant/interesting

Subgroup search: Having found a setting where X is quasi-randomly assigned, try various subgroups (e.g. young Asian men) until find a significant relationship

Which of these is better with "credibility revolution"?
Which is worse?

Some problems with NHST (cont'd)

You reason: "If my theory is correct, then X should be positively related to Y."

You set up hypothesis test:

Null hypothesis: X is not related to Y

Alternative hypothesis: X is positively related to Y

Your result: Null hypothesis rejected.

Is your theory therefore correct?

Some problems with NHST (cont'd)

You reason: "If my theory is correct, then X should be positively related to Y."

You set up hypothesis test:

Null hypothesis: X is not related to Y

Alternative hypothesis: X is positively related to Y

Because the world is complicated, the null hypothesis is definitely false (unless X or Y is under your control and random).

So why test it?

Two ways forward

Standalone "well-identified" studies

Rejection of null leaves little doubt that "theory" is correct: if X and Y are related, it is because X affects Y. (Ceteris paribus likely to hold.)

Body of evidence, multiple risky tests

In any particular study, rejection of null could have many interpretations, but collectively studies point toward "theory" being correct.

Why most social science puzzles aren't puzzling and many social science findings inconclusive

Our theories are weak and the world is complicated.

Theories are either

- very flexible: rational choice, "ideas, interests, and institutions", realism, constructivism, Freudianism, or
- very partial (i.e. not attempting to predict or describe reality): Downsian competition, rationalist explanations of war

Our theories do not lead to *critical tests* (like Eddington's eclipse test, 1919). Social science is not physics.

Please remember this advice if your research involves any claims about effects

Suppose there were no constraints (time, money, ethics, the number of countries). What is the most informative experiment I could run to measure the effect I want to study?

Benefits: Clarifies to you (and reader)

- what you are trying to study
- what challenges you face
- what feasible designs actually exist

Some big questions about design-based inference and the "credibility revolution"

- Does a study on elections in French villages tell us anything about national elections? (external validity)
- What about explanation? What does the study in French villages tell us about why turnout is higher in PR countries (the puzzle to be explained)?
- What about "theory-testing"? What theory is tested when the setting for our analysis was carefully chosen?
- What about "effects of causes" questions that can't be answered this way: what is the effect of globalization?

Replication movement and DA-RT

http://www.dartstatement.org/

Petition to delay DA-RT implementation

Petition to Delay DA-RT Implementation

November 3, 2015 [list includes those who signed of November 8 5:15 pm EST]]

Dear Colleagues,

We write as concerned members of the American Political Science Association to urge an important amendment to the statement, "Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors." In the joint statement, dated October 6, 2014, journal editors committed their respective journals to a set of principles, to be implemented by January 15, 2016.

DA-RT organizers have made many efforts over the past five years to reach out to members of the profession through various symposia and meetings. However, these issues began to gain widespread attention only when the journal editors signed the statement of October 6, 2014 and panels at the 2015 annual meeting of the American Political Science Association brought the issue to the attention of many scholars who had not realized the possible implications of that statement for their own research, despite the previous outreach activities. Conversations at the panels, roundtables, section business meetings, and other venues at the recent annual meeting demonstrated that members of the Association have only just begun to grapple with the

Pre-registration movement and EGAP