Applied Microeconometrics I

Lecture 3: Randomized controlled trials (continued)

Stefano Lombardi

Aalto University

September 12, 2023

Lecture Slides

What did we do last time?

- Correlation does not imply causation: $Corr(x, y) \neq 0$ is consistent with:
  1. x causes y
  2. y causes x
  3. z causes x and y
- Experiments in physical sciences (the scientific solution)
  1. Temporal stability
  2. Causal transience
  3. Unit homogeneity
- Why are these assumptions unlikely to hold in social sciences?
- Economists rely on a statistical solution

What did we do last time?

- Define treatment variable $D_i = \{0, 1\}$
- Potential outcomes for each i:

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

- Causal effect for i:

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

- but we only observe:

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$$

What did we do last time?

- The fundamental problem of causal inference: We can never observe: $Y_{1i} - Y_{0i}$
- We would instead want to know the average treatment effect on the treated (ATET): $E[Y_{1i} - Y_{0i} | D_i = 1]$
- But we only observe:

$$E[Y_i | D_i = 1] - E[Y_i | D_i = 0] \quad = \quad ATET + \text{selection bias}$$

What did we do last time?

- The selection bias is 0 only if $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0]$. Is this plausible?
- ATET is in general different from ATE (depending on the application, one of the two might be more relevant)
- ATE is equal to ATET only if the average treatment effect is constant in the population

What did we do last time?

- **Statistical solution**: assign $D_i = 1$ randomly
- As a result $D_i = 1$ is independent of all individual attributes and of both Y_{1i} and Y_{0i}
- Then:

$$\begin{aligned} E(Y_{1i}|D_i = 1) &= E(Y_{1i}|D_i = 0) \\ E(Y_{0i}|D_i = 1) &= E(Y_{0i}|D_i = 0) \end{aligned}$$

- and it holds that:

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E(Y_{1i} - Y_{0i}) \end{aligned}$$

- Numerical example and Stata example

Today: RCTs (continued)

1. Which questions (in principle) can have causal answers?
2. Several important points about RCT's
3. Potential drawbacks of RCTs
4. Examples of RCT's
   - Tennessee STAR experiment
   - Electoral Fraud in Russia
   - Other examples (optional)

Today: RCTs (continued)

- R-tutorial tomorrow given by Ramin Izadi
- Download and install R-studio
- Cristina Bratu will teach the next two lectures

Which questions have causal answers?

- Thinking in terms of counterfactual outcomes and (possibly imaginary) randomized trials also helps one to formulate research questions more precisely
- Think of the following statements due to Holland (1986): She did well on the exam because...
  - A. she is a woman
  - B. she studied for it
  - C. she was coached by her teacher
- In the case of each statement we should:
  - Think about what is the *cause* according to the statement
  - Can we manipulate it *ceteris paribus*?

Which questions have causal answers?

- Statement C is the most straightforward. The cause is coaching and we can easily think of manipulating it randomly.
- In statement A the cause is gender. Could this be manipulated (and therefore be a cause)? Or should we re-frame the question?
- Statement B is typically encountered in economics:
  - What is the cause in this statement?
  - Can we manipulate it in an experiment?
- Thinking about these issues is helpful in defining research questions

Several important points about RCT's

1. Role of theory
2. Ideal experiment
3. *Fundamentally unidentified* questions
4. Consistency

Randomized controlled trials

Role of theory

- Theory can be helpful in the interpretation of the results
- Example: Zinovyeva and Bagues (2015)

Randomized controlled trials

Role of theory: Zinovyeva and Bagues (2015)

What is the effect of connections on academic promotions?

- In Spain, promotion decisions are taken by a committee of professors (which might, by chance, know the candidate).
- Committee composition and academic promotions: does it help to have your supervisor to evaluate you?
- Random allocation of committees in Spanish academia
- The effect of connections can work through:
  - "Bias" (e.g., favoritism)
  - Information (evaluator knows more than what's in the CV)
- The effect of connections in a RCT is identified, but is a combination of these two forces. *How to disentangle them?*

Randomized controlled trials

Role of theory: Zinovyeva and Bagues (2015)

- Important to know the mechanism (policy implications)
- Derive predictions from a theoretical model:
    - "Biased" promotions lead to worse productivity outcomes
    - Informed promotions lead to better productivity outcomes
- Define productivity via publications (good proxy)
- Estimate the causal effect of connections on productivity of the promoted candidates via RCT
- *Ex-ante*, the theoretical framework predicts that if "bias" dominates, the connection effect sign is negative

Randomized controlled trials

Ideal experiment

- Ideal experiment helps to formulate causal question precisely
- Example: **discrimination** in the hiring process is relevant to understand the functioning of labor markets.
    - Discrimination is multidimensional problem. What can we conclude if we observe differences in outcomes across groups?
    - Problem: if attributes cannot be manipulated/randomized (e.g., race, gender, age), then they cannot be causes.
- Manipulate *perception* of attributes in hiring process:
    - Goldin and Rouse (2000): gender (symphony orchestras)
    - Bertrand and Mullainathan (2003): race (fictitious CV's)

Bertrand and Mullainathan (2003)

- Discrimination used to be studied with audit studies where experimental candidates were sent to job interviews
- Problems with this approach:
  1. Very small samples (very expensive)
  2. Impossibility of conducting double-blind studies
  3. Artificiality of the setting (external validity?)
- Bertrand and Mullainathan: Apply for jobs by sending CV's. Manipulate perceptions of race by using distinctively ethnic names (otherwise CV information identical). Are callback rates lower for individuals with "black-sounding" names?
- Callback rates are lower for black-sounding names
- Black names benefit less from CV enhancements than white

Mean callback rates by name types

Bertrand and Mullainathan (2003)

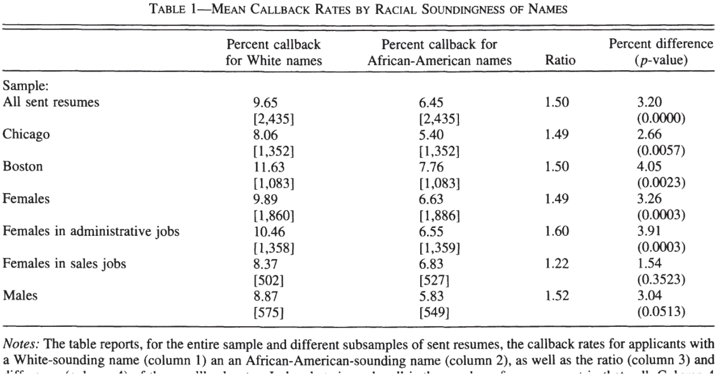


TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

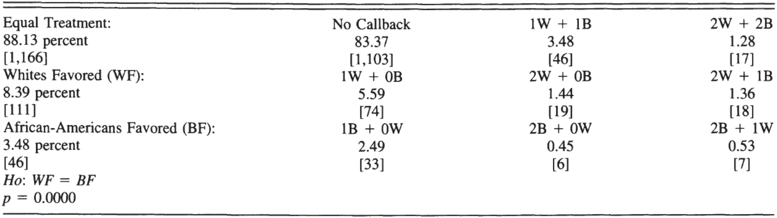
	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (*p*-value)
Sample:				
All sent resumes	9.65 [2,435]	6.45 [2,435]	1.50	3.20 (0.0000)
Chicago	8.06 [1,352]	5.40 [1,352]	1.49	2.66 (0.0057)
Boston	11.63 [1,083]	7.76 [1,083]	1.50	4.05 (0.0023)
Females	9.89 [1,860]	6.63 [1,886]	1.49	3.26 (0.0003)
Females in administrative jobs	10.46 [1,358]	6.55 [1,359]	1.60	3.91 (0.0003)
Females in sales jobs	8.37 [502]	6.83 [527]	1.22	1.54 (0.3523)
Males	8.87 [575]	5.83 [549]	1.52	3.04 (0.0513)

Notes: The table reports, for the entire sample and different subsamples of sent resumes, the callback rates for applicants with a White-sounding name (column 1) an an African-American-sounding name (column 2), as well as the ratio (column 3) and

Distribution of callbacks

Bertrand and Mullainathan (2003)

TABLE 2—DISTRIBUTION OF CALLBACKS BY EMPLOYMENT AD



Equal Treatment:	No Callback	1W + 1B	2W + 2B
88.13 percent	83.37	3.48	1.28
[1,166]	[1,103]	[46]	[17]
Whites Favored (WF):	1W + 0B	2W + 0B	2W + 1B
8.39 percent	5.59	1.44	1.36
[111]	[74]	[19]	[18]
African-Americans Favored (BF):	1B + 0W	2B + 0W	2B + 1W
3.48 percent	2.49	0.45	0.53
[46]	[33]	[6]	[7]
Ho: *WF* = *BF*			
p = 0.0000			

Mean callback rates and CV quality

Bertrand and Mullainathan (2003)

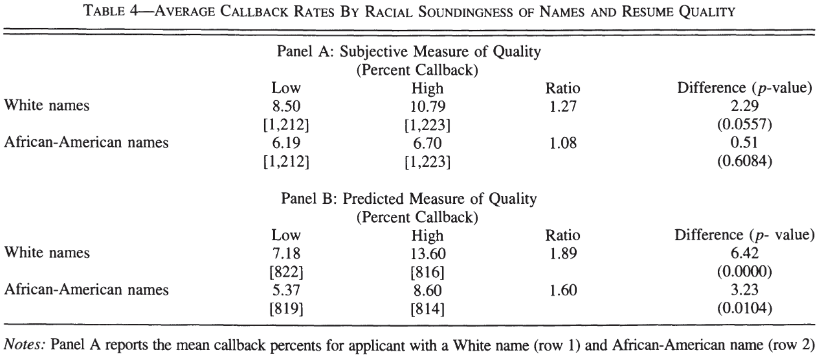


TABLE 4—AVERAGE CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES AND RESUME QUALITY

	Panel A: Subjective Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (*p*-value)
White names	8.50	10.79	1.27	2.29
	[1,212]	[1,223]		(0.0557)
African-American names	6.19	6.70	1.08	0.51
	[1,212]	[1,223]		(0.6084)
	Panel B: Predicted Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (*p*- value)
White names	7.18	13.60	1.89	6.42
	[822]	[816]		(0.0000)
African-American names	5.37	8.60	1.60	3.23
	[819]	[814]		(0.0104)

Notes: Panel A reports the mean callback percents for applicant with a White name (row 1) and African-American name (row 2)

The effect of CV quality on callbacks

Bertrand and Mullainathan (2003)

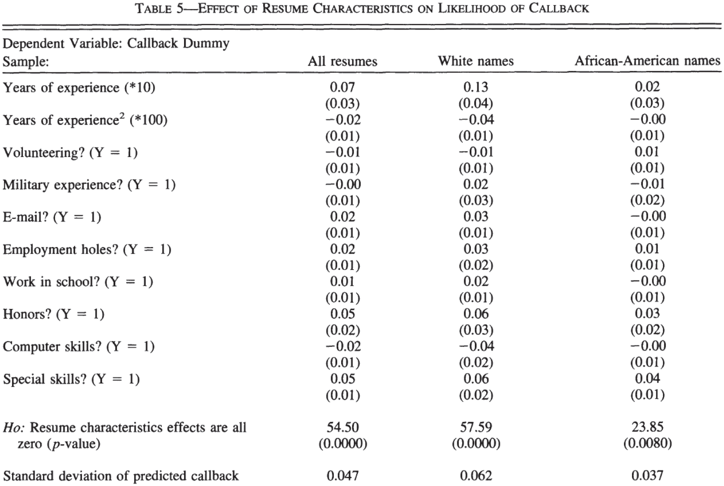


TABLE 5—EFFECT OF RESUME CHARACTERISTICS ON LIKELIHOOD OF CALLBACK

Dependent Variable: Callback Dummy Sample:	All resumes	White names	African-American names
Years of experience (*10)	0.07	0.13	0.02
	(0.03)	(0.04)	(0.03)
Years of experience^2 (*100)	−0.02	−0.04	−0.00
	(0.01)	(0.01)	(0.01)
Volunteering? (Y = 1)	−0.01	−0.01	0.01
	(0.01)	(0.01)	(0.01)
Military experience? (Y = 1)	−0.00	0.02	−0.01
	(0.01)	(0.03)	(0.02)
E-mail? (Y = 1)	0.02	0.03	−0.00
	(0.01)	(0.01)	(0.01)
Employment holes? (Y = 1)	0.02	0.03	0.01
	(0.01)	(0.02)	(0.01)
Work in school? (Y = 1)	0.01	0.02	−0.00
	(0.01)	(0.01)	(0.01)
Honors? (Y = 1)	0.05	0.06	0.03
	(0.02)	(0.03)	(0.02)
Computer skills? (Y = 1)	−0.02	−0.04	−0.00
	(0.01)	(0.02)	(0.01)
Special skills? (Y = 1)	0.05	0.06	0.04
	(0.01)	(0.02)	(0.01)
Ho: Resume characteristics effects are all zero (*p*-value)	54.50 (0.0000)	57.59 (0.0000)	23.85 (0.0080)
Standard deviation of predicted callback	0.047	0.062	0.037

Discussion

Bertrand and Mullainathan (2003)

- What is the ideal experiment here?
- What is the cause that is manipulated in this experiment?
- Limitations of the experiment
  - Does this experiment answer the question that the authors are interested in? Which type of discrimination can we study?
  - Outcome variable
  - Representativeness of the names
  - How powerful is the treatment?

Randomized controlled trials

Fundamentally unidentified questions

- *No causation without (in principle) manipulation*
- Manipulation defines the causal answer we get
- Questions that cannot be answered by any experiment are *fundamentally unidentified* (ill-defined questions)
    - Bertrand and Mullainathan (2003): Can we manipulate race?
    - The effect of start age on first grade test scores
        - start age = age - time in school

Randomized controlled trials

Consistency

- ATET is estimated by comparing outcome means
  - OLS, regressing outcome on treatment dummy
  - Randomization ensures that *estimated* effect is consistent for the *true* population ATET.
- Crucial assumption
  - The treatment was assigned randomly (D_i is unrelated to any relevant variables that are correlated with the outcome)
- Randomization checks
  - How was randomization conducted? Any room for manipulation?
  - Are the covariates balanced across treatment groups *before* the treatment was assigned? (pre-determined covariates)
  - Similarly, do results change when you add covariates?
- Should we care about R-squared for consistency?

Potential drawbacks of RCTs

Experiments provide a simple strategy to solve the *selection bias*; however, they can have a number of **potential problems**:

1. **Implementation issues**
    - Compliance (e.g., take the treatment even if in control group)
    - Attrition (people leave treat. or control groups)
    - Cost, political issues
2. **Ethical issues**: why don't treat everyone?
    - The ethical argument is not obvious when (i) the treatment cannot be applied to everybody (budget constraints); (ii) the optimal assignment rule is unknown; (iii) randomization is fair.
3. **Hawthorne effect**
    - The fact that people know about being part of an experiment make them behave differently (external validity?)
    - Landsberger (1950); Levitt, List (2011); Behaghel et al. (2015)

Potential drawbacks of RCTs

4. **General equilibrium (GE) issues**
    - Randomization requires that D_i does not affect i's potential outcomes, but also those of *other* units
    - Two problems of GE/spillover effects:
        - GE effects invalidate results of RCT's
        - We are often interested in treatments where the effect of treatment depends on *how many* individuals receive the treatment (general equilibrium, spillover effects)
    - Can we account for (and study!) GE effects in RCTs?
    - **Example**: Crepon et al. (2012) ▸
5. **External validity vs Internal validity**
    - Problem also with other identification strategies
    - Structural models

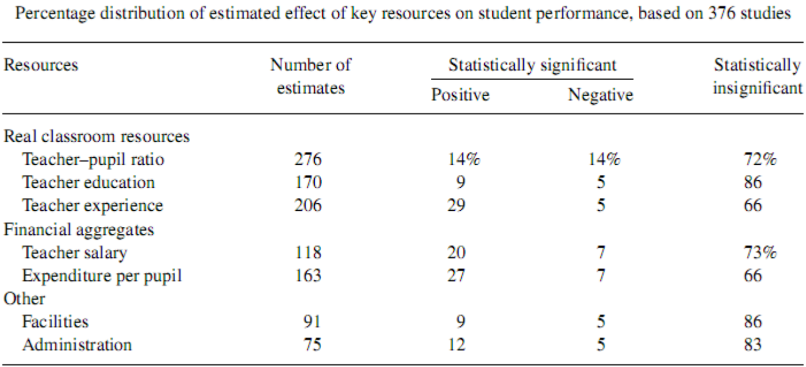
Examples of RCTs

- Tenessee STAR experiment
- Electoral fraud in Russia
- Other examples (optional)

Example: Does class size affect students' performance?

Tennessee STAR experiment

- How can we improve students' performance?
- Should we devote more resources to reduce class size?
  - Example: Should we split this course in two separate groups?
- A large number of observational studies tend to find that class size is not generally associated with better student performance
  - Hanushek (1997): "No strong or systematic relationship between school inputs and student achievement"
- We can explain the (spurious correlation) between class size and students' performance with rational choices of principals
  - How to solve the selection bias problem?



Percentage distribution of estimated effect of key resources on student performance, based on 376 studies

Resources	Number of estimates	Statistically significant		Statistically insignificant
		Positive	Negative	
Real classroom resources				
Teacher–pupil ratio	276	14%	14%	72%
Teacher education	170	9	5	86
Teacher experience	206	29	5	66
Financial aggregates				
Teacher salary	118	20	7	73%
Expenditure per pupil	163	27	7	66
Other				
Facilities	91	9	5	86
Administration	75	12	5	83

Example: Does class size affect students' performance?

Tennessee STAR experiment

Tennessee STAR experiment

- Cost: $12 million
- A cohort of kindergartners in 1985/86: 11,600 children in 80 schools
- The study ran for four years
- Three treatments:
  1. small classes with 13-17 children
  2. regular classes with 22-25 children without a teacher's aide.
  3. regular classes with 22-25 children with a teacher's aide.
- *Within each school*, students are randomly assigned to one of these groups

Example: Does class size affect students' performance?

Tennessee STAR experiment

Krueger (1999) analyzes the short-run effects of the experiment.

Main findings:

1. performance on standardized tests increases by four percentile points the first year students attend small classes
2. the test score advantage of students in small classes expands by about one percentile point per year in subsequent years
3. teacher aides and measured characteristics have little effect
4. larger effect for minority students/on free lunch

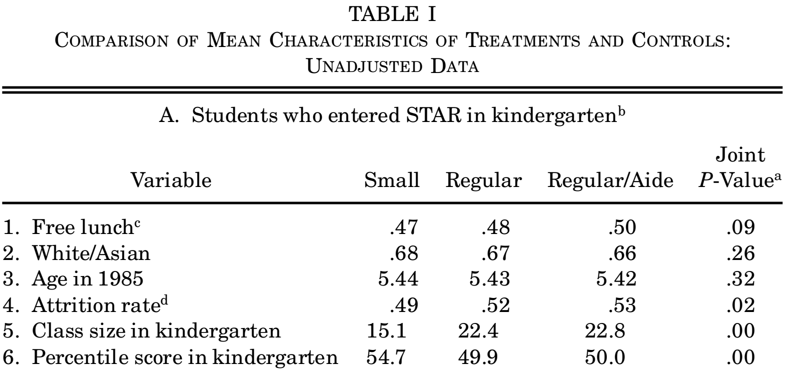


TABLE I

Comparison of Mean Characteristics of Treatments and Controls: Unadjusted Data

A. Students who entered STAR in kindergarten[b]

Variable	Small	Regular	Regular/Aide	Joint *P*-Value[a]
1. Free lunch[c]	.47	.48	.50	.09
2. White/Asian	.68	.67	.66	.26
3. Age in 1985	5.44	5.43	5.42	.32
4. Attrition rate[d]	.49	.52	.53	.02
5. Class size in kindergarten	15.1	22.4	22.8	.00
6. Percentile score in kindergarten	54.7	49.9	50.0	.00

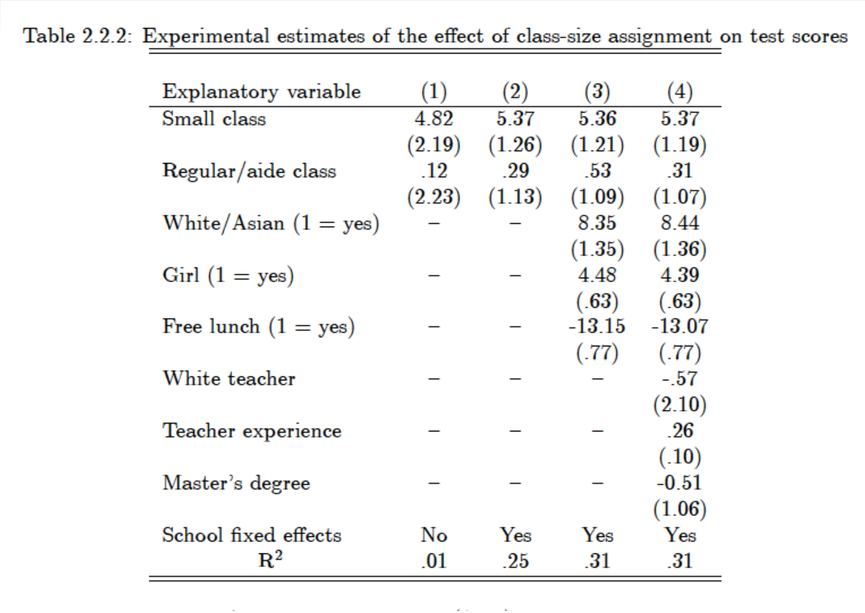


Table 2.2.2: Experimental estimates of the effect of class-size assignment on test scores

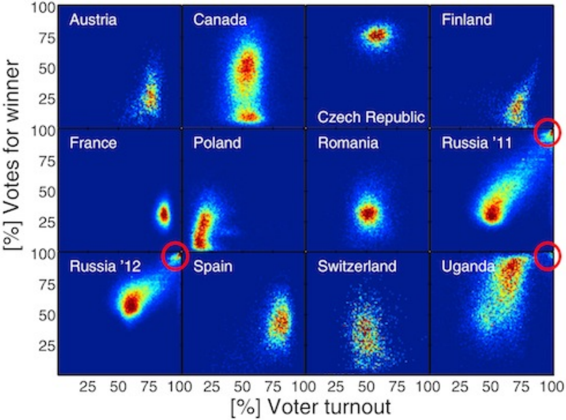
Explanatory variable	(1)	(2)	(3)	(4)
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes)	–	–	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	–	–	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	–	–	-13.15	-13.07
			(.77)	(.77)
White teacher	–	–	–	-.57
				(2.10)
Teacher experience	–	–	–	.26
				(.10)
Master's degree	–	–	–	-0.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Electoral fraud in Russia

- Motivation:
  - Is there any electoral fraud is Russia?
  - How much?
- Available evidence:
  - Anecdotal evidence
  - Statistical evidence

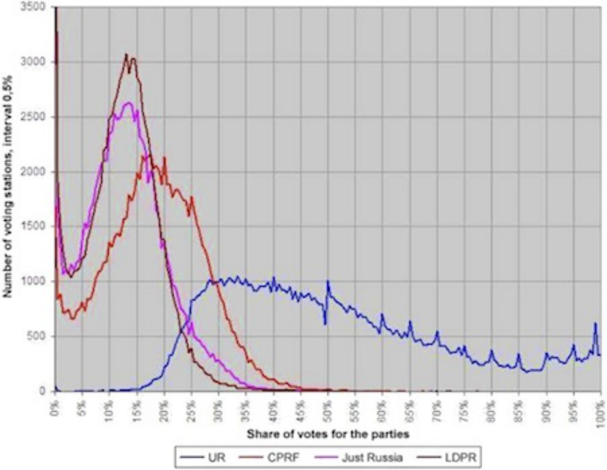
Circumstancial evidence (i)

Bimodal distribution of votes



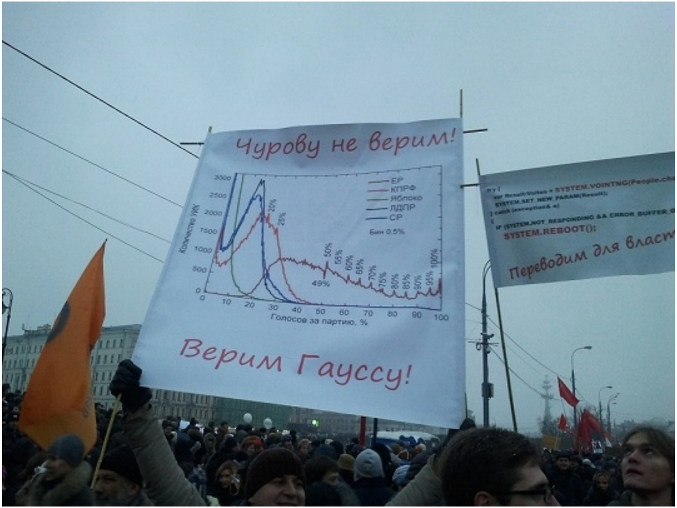
Circumstancial evidence (ii)

Spikes in the distribution of votes for United Russia
Kobak, Shpilkin and Pschenichnikov (2016)



Circumstancial evidence (ii)

"We do not believe Churov [the head of the electoral committee], we believe Gauss!"



```
try {
  var Result-Voites = SYSTEM.VOINTNG(People.cho
  SYSTEM.SET_NEW_PARAM(Result);
} catch (exception& e)
{
  if (SYSTEM.NOT_RESPONDING && ERROR_BUFFER_O
    SYSTEM.REBOOT();
```

Electoral fraud in Russia

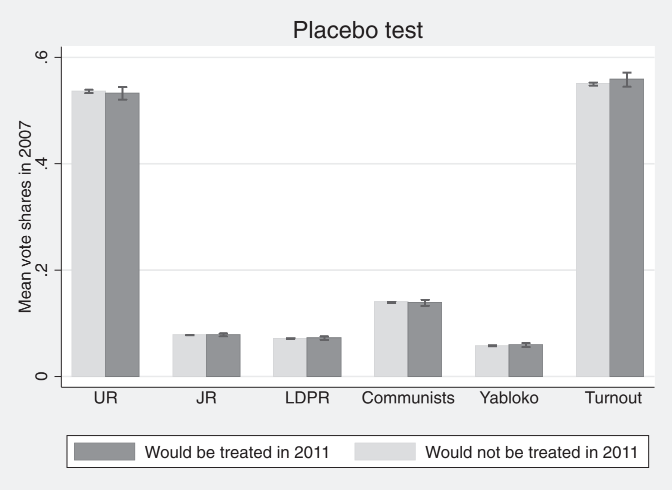
Rephrase slightly our question in a *treatment effects* fashion:

- Would electoral results change if there were independent observers in the polling stations?
- To address this question, we can try to send observers to some (non-randomly selected) polling stations, as some NGOs and international organizations do.
    - How informative would this be?
- Can you propose a better approach?
    - Enikolopov, Korovkina, Petrova, Sonin and Zakharov (2013)

Field experiment estimate of electoral fraud in Russian parliamentary elections

Enikolopov, Korovkin, Petrova, Sonin and Zakharov (2013)

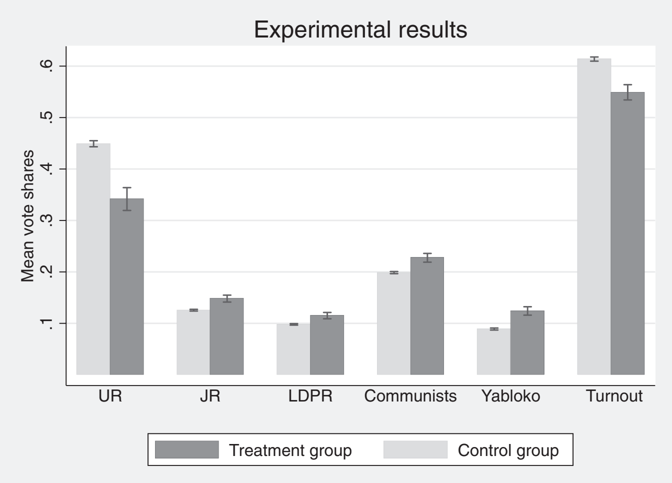
- Random assignment of independent observers to 156 of 3,164 polling stations in the city of Moscow
- Within each district, polling stations were sorted according to their official number assigned by Central Election Committee. Every 25th polling station within an electoral district, starting from the first, was assigned for observation


Placebo test
Mean vote shares in 2007
0
.2
.4
.6
UR
JR
LDPR
Communists
Yabloko
Turnout
Would be treated in 2011
Would not be treated in 2011

Field experiment estimate of electoral fraud in Russian parliamentary elections

Enikolopov, Korovkin, Petrova, Sonin and Zakharov (2013)

- Treatment:
  - Observers can only prevent the most obvious types of fraud
  - Not full compliance: Some of these observers were removed before the vote counting process was finished


Experimental results
Mean vote shares
.1
.2
.3
.4
.5
.6
UR
JR
LDPR
Communists
Yabloko
Turnout
Treatment group
Control group

Field experiment estimate of electoral fraud in Russian parliamentary elections

Enikolopov, Korovkin, Petrova, Sonin and Zakharov (2013)

- Main results
  - The actual share of votes for the incumbent United Russia party is 11 percentage points lower in treatment areas (36% instead of 47%).
  - The turnout at the polling stations with observers was lower by 6.5 percentage points
- Interpretation?

More examples (optional)

Which **experiment** could be used to capture the causal effect?

- Would it be profitable for the call center of a travel agency to allow their employees to work from home?
  - Bloom et al 2012
- Are employees more satisfied if they are informed about the salaries of their colleagues?
  - Card et al 2011
- Does the gender composition of hiring committees matter?
  - Bagues and Esteve-Volart 2010
- Do monetary incentives crowd out intrinsic motivation?
  - Gneezy and Rustichini 2000
  - Lacetera et al. 2012

More examples (optional)

Which **experiment** could be used to capture the causal effect?

- Do “modern managerial” practices increase firms’ productivity? (lean manufacturing principles)
  - Bloom et al. 2010
- An increase in the salaries offered in the public sector attracts candidates that are less committed to public service
  - Dal Bo, Finan and Rossi 2012
- How can we decrease impact of AIDS in Subsaharian Africa?
  - Dupas 2011
- Do Indian teachers react to incentives?
  - Duflo et al 2012

General equilibrium effects in RCT's

- Crepon et al (2012) (Active Labor Market Policies)
- **Question**: Does job search assistance affect employment prospects of unemployed job-seekers?
- If the effect is large, do treated job-seekers crowd out non-treated job-seekers? (this invalidates standard RCT's)
- **Solution** is new type of RCT, *double randomization*:
  1. Assign to each local job market the *share of treated* job-seekers randomly (e.g., 5%, 20% with probabilities 0.5)
  2. Within each local job market, assign the JSA *treatment* to job-seekers randomly (according to the share previously drawn)
- Estimate the effect of assignment to treatment on the treated (within labor market)
- Estimate the effect of share assigned to treatment on the controls (exploiting share variation *across* job markets)