



BERKELEY INITIATIVE FOR TRANSPARENCY  
IN THE SOCIAL SCIENCES

# Data Sharing and Citations

## Causal Evidence

Garret Christensen<sup>1</sup>   Allan Dafoe<sup>2</sup>   Edward Miguel<sup>3</sup>

<sup>1</sup>Berkeley Institute for Data Science, UC Berkeley

<sup>2</sup>Department of Political Science, Yale University

<sup>3</sup>Department of Economics, UC Berkeley

August 31, 2017 BIDS

PRELIMINARY—Please do not cite.

Thanks to David Birke, Terri Cruz, Maxim Guzman, Baiyue Cao, Evey Huang, Rachel Kim, Ravina Pattni, Kevin Khuu

# Outline

- 1 Introduction
- 2 Methods
- 3 Results
- 4 Conclusion

Slides available at:

<http://www.github.com/bitss/citations>

# Data Sharing Incentives

- Shared data is a public good. (See Newton 1675)
- Public goods are often undersupplied.
- Is there private incentive?
  - Promotion & tenure
  - Citations

# Citations

- Signal of quality (?)
- Facilitates other researchers building off your work



Angrist Data Archive

Angrist, Jordà and Kuersteiner (2016)

Angrist, Oreopoulos and Williams (2014)

Angrist and Fernandez-Val (2010)

Angrist and Kuersteiner (2010)

Angrist and Lavy (2009)

Angrist, Lang, and Oreopoulos (2009)

Angrist and Kugler (2008)

Angrist, Chin, and Godoy (2008)

Angrist (2006)

Angrist, Bettinger, and Kremer (2006)

Angrist, Chernozhukov, and Fernandez-Val (2006)

Angrist and Lang (2004)

Angrist and Kugler (2003)

Abadie, Angrist, and Imbens (2002)

Angrist (2002)

# Angrist Data Archive

## Data and programs from Angrist journal articles

Follow these links to data sets and programs from a number of my papers. In some cases, I've taken advantage of the opportunity to make minor corrections. Some old SAS programs have been converted to Stata (but SAS is good for you, so I've left some that way). Feel free to use these files for teaching or research (with attribution!).

Many of these replication files are also available in my [IQSS Dataverse](#); use the Dataverse for online variable selection, automatic subsetting, and to download files in alternative formats.

DATA NEWS: June 2017

We've published in JBES. Latest paper has been posted: [Angrist, Jordà and Kuersteiner \(2016\)](#).

DATA NEWS: February 2016

We've posted data for [Angrist, Oreopoulos and Williams \(2014\)](#)

DATA NEWS: NOVEMBER 2015

We've posted data for [Angrist and Lavy \(2001\)](#).

DATA NEWS: DECEMBER 2014

I've uploaded a do file to replicate Tables 1 and 2 in [Angrist and Fernandez-Val \(2010\)](#). For Tables 3 and 4, see [Ivan Fernandez-Val's data archive](#).

DATA NEWS: JULY 2011

I've uploaded data and programs to replicate the cross-district analysis of Metco in [Angrist and Lang \(2004\)](#).

DATA NEWS: FEBRUARY 2011

Adriana Kugler and I have posted the Colombian rural household survey data for our [Rural Windfall](#) paper (May 2008 ReStat). Download 'em quick, before we get busted!

DATA NEWS: FEBRUARY 2009

Data sets and programs from many of the papers by other authors referenced in [Mostly Harmless Econometrics](#) are posted in the [MHE Data Archive](#).

DATA NEWS: JUNE 2008

The posting for Angrist and Krueger (1991) (below, but not in the dataverse) now includes all of the 1970 and 1980 census cohorts used in the paper, including all covariates. The smaller 1980 census extract for men born 1930-39 is also still available as an ASCII file.

## Articles

[\[book\] Mostly harmless econometrics: An empiricist's companion](#)

[JD Angrist, JS Pischke](#) - 2008 - [books.google.com](#)

## Case law

The core methods in today's econometric toolkit are linear regression for statistical control, instrumental variables methods for the analysis of natural experiments, and differences-in-differences methods that exploit policy changes. In the modern experimentalist paradigm, Cited by 8629 Related articles All 12 versions Cite Save More

## My library

## Any time

Since 2017

Since 2016

Since 2013

Custom range...

[Identification of causal effects using instrumental variables](#)

[JD Angrist, GW Imbens, DB Rubin](#) - Journal of the American ..., 1996 - Taylor & Francis

Abstract We outline a framework for causal inference in settings where assignment to a binary treatment is ignorable, but compliance with the assignment is not perfect so that the receipt of treatment is nonignorable. To address the problems associated with comparing Cited by 4269 Related articles All 30 versions Cite Save More

[\[PDF\] zmjones.com](#)

## Sort by relevance

## Sort by date

[Identification and estimation of local average treatment effects](#)

[J Angrist, G Imbens](#) - 1995 - [nber.org](#)

We investigate conditions sufficient for identification of average treatment effects using instrumental variables. First we show that the existence of valid instruments is not sufficient to identify any meaningful average treatment effect. We then establish that the combination Cited by 3619 Related articles All 22 versions Cite Save More

[\[PDF\] baylor.edu](#)

[Does compulsory school attendance affect schooling and earnings?](#)

[JD Angrist, AB Keueger](#) - The Quarterly Journal of Economics, 1991 - [academic.oup.com](#)

Abstract We establish that season of birth is related to educational attainment because of school start age policy and compulsory school attendance laws. Individuals born in the beginning of the year start school at an older age, and can therefore drop out after Cited by 2343 Related articles All 23 versions Cite Save More

[\[PDF\] princeton.edu](#)

[Instrumental variables and the search for identification: From supply and demand to natural experiments](#)

[J Angrist, AB Krueger](#) - 2001 - [nber.org](#)

ABSTRACT The method of instrumental variables was first used in the 1920s to estimate supply and demand elasticities, and later used to correct for measurement error in single-equation models. Recently, instrumental variables have been widely used to reduce bias Cited by 1913 Related articles All 50 versions Cite Save More

[\[PDF\] mit.edu](#)

[Using Maimonides' rule to estimate the effect of class size on scholastic achievement](#)

[\[PDF\] ens.fr](#)

☐ include patents

☒ include citations

☒ Create alert

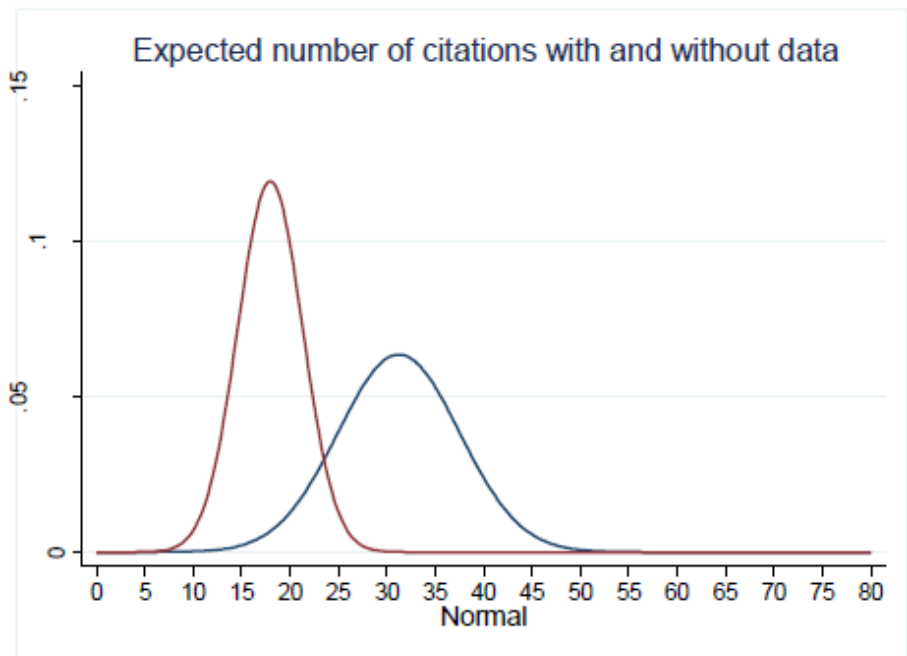


# Existing Evidence

- Piwowar, Day, Fridsma (2007): 69% more citations for cancer microarray clinical trials papers (N=85).
- Piwowar, Vision (2013): 9% more citations for gene expression microarray papers with public data (N=10,555).
- *Journal of Peace Research*
  - Yes: Gleditsch, Metelits, Strand. 2003. "Posting Your Data: Will You Be Scooped or Will You Be Famous?"
  - No: Abbott 2007
  - Yes: Strand, Nordkvelle, Gleditsch. 2014. "Posting Your Data: Will You Remain Famous?"



Figure 7: Predicted citations for an average article w/quant. data analysis in JPR



# The Case of Political Science

Exploit plausibly exogenous variation in data availability caused by the abrupt change in editorial policy at a top political science journal, *The American Journal of Political Science* (AJPS).

Rick Wilson became the editor on January 1, 2010:

*“If a manuscript is accepted for publication it will not be published unless the first footnote explicitly notes where the data used in the study can be obtained for purposes of replication and should note any sources that funded the research.”*

# The Case of Political Science

The first issue Wilson edited was published in October 2010. After discussion with the board members in April of 2012, the policy was expanded to require posting data in the journal's public archive at Harvard's Dataverse and Wilson strengthened his enforcement of this policy. This policy was printed in the July 2012 issue, and was enforced thereafter.

There was no policy change at the other top political science journal, *American Political Science Review*, (APSR).

# The Case of Economics\*

- The Journal of Money, Credit, and Banking Project (Dewald, Thursby, Anderson 1986)
- Verifying the Solution from a Nonlinear Solver: A Case Study (McCullough, Vinod 2003)
- Ben Bernanke made the *American Economic Review* policy mandatory in 2005.
- *Quarterly Journal of Economics* didn't require data sharing until 2016.

# Replication in Empirical Economics: *The Journal of Money, Credit and Banking Project*

By WILLIAM G. DEWALD, JERRY G. THURSBY, AND RICHARD G. ANDERSON\*

*This paper examines the role of replication in empirical economic research. It presents the findings of a two-year study that collected programs and data from authors and attempted to replicate their published results. Our research provides new and important information about the extent and causes of failures to replicate published results in economics. Our findings suggest that inadvertent errors in published empirical articles are a commonplace rather than a rare occurrence.*

# Pre-Analysis Plan

- Short pre-analysis plan before data collection.
- Available at <https://osf.io/qxpr6/>
- Use 2SLS to estimate causal effect (LATE)
  - Angrist, Imbens, Ruben (1993)
  - Angrist, Imbens (1994)



# 2SLS in 60 Seconds or Less

$$Y_i = \alpha + \beta \cdot X_i + u_i$$

Problem: X of interest is endogenous. (Omitted variable bias—some unobservable is correlated with Y and X of interest.)

Solution: Find an instrument.

$$\text{corr}(Z, u) = 0$$

$$\text{corr}(Z, X) \neq 0$$

Biased, but consistent. (Get large N!)

## 2SLS in More Detail

### Implementation:

- 1 Predict X with Z. (First Stage)
- 2 Regress Y on predicted X's. (Second Stage)
- 3 Adjust standard errors.

### Assumptions:

- Relevance. (Strong first stage,  $\text{corr}(Z, X) \neq 0$ ;  $F > 10$ )
- Exclusion Restriction. ( $\text{corr}(Z, u) = 0$ ; Instrument only effects outcome through included X's.)

## 2SLS Examples

- Best: Randomized Trial, incomplete adoption, use treatment assignment as instrument for treatment, get TOT and ITT. (Wald Estimator)
- Vietnam draft lottery predicts service, get unbiased effect of service on earnings.
- Quarter of Birth as instrument for schooling, get effect of schooling on earnings (with no ability bias).
- Settler mortality as instrument for institutions, get effect of institutions on GDP.
- Rainfall shocks predict GDP, get effect of GDP on civil conflict.

# Our Estimation Strategy

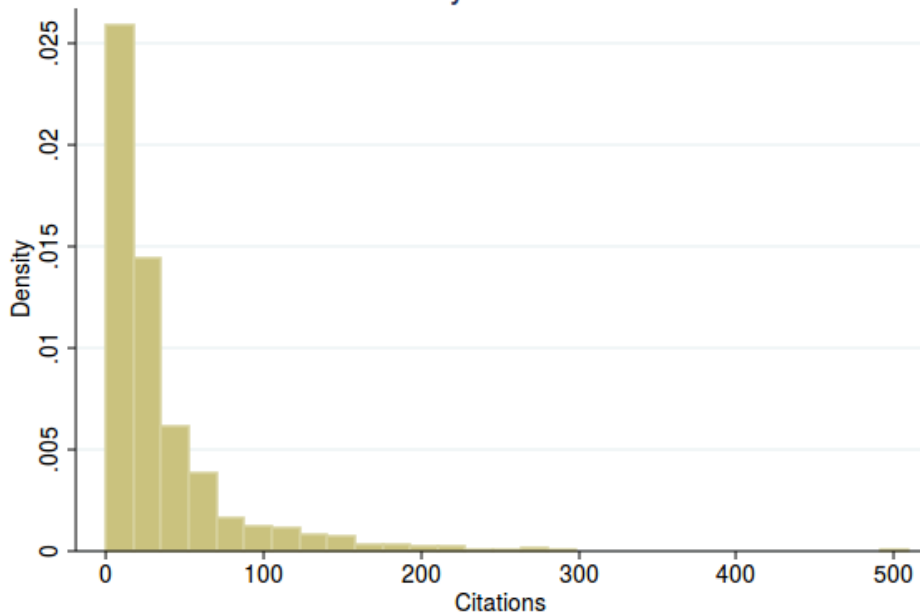
$$\begin{aligned}
 \text{availability}_i = & \alpha_1 + \beta_1 \text{AJPS}_i + \beta_2 \text{Post2010}_i + \beta_3 \text{Post2012}_i \quad (1) \\
 & + \beta_4 \text{AJPS} * \text{Post2010}_i + \beta_5 \text{AJPS} * \text{Post2012}_i \\
 & + g_1(\text{Time}) + h_1(\text{Year}) + \nu_i
 \end{aligned}$$

$$\begin{aligned}
 \text{citations}_i = & \alpha_2 + \eta_1 \text{AJPS}_i + \eta_2 \text{Post2010}_i + \eta_3 \text{Post2012}_i \quad (2) \\
 & + \eta_4 \hat{\text{availability}}_i + g_2(\text{Time}) + h_2(\text{Year}) + u_i
 \end{aligned}$$

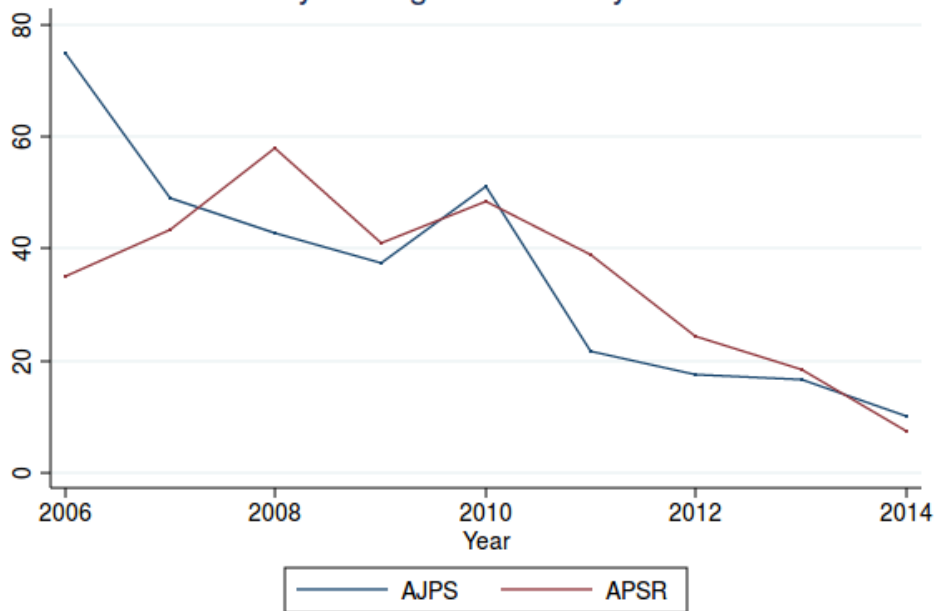
# Summary Statistics

- Citations highly concentrated.
- Citations increase over time.
- Citations affected by journal policy.

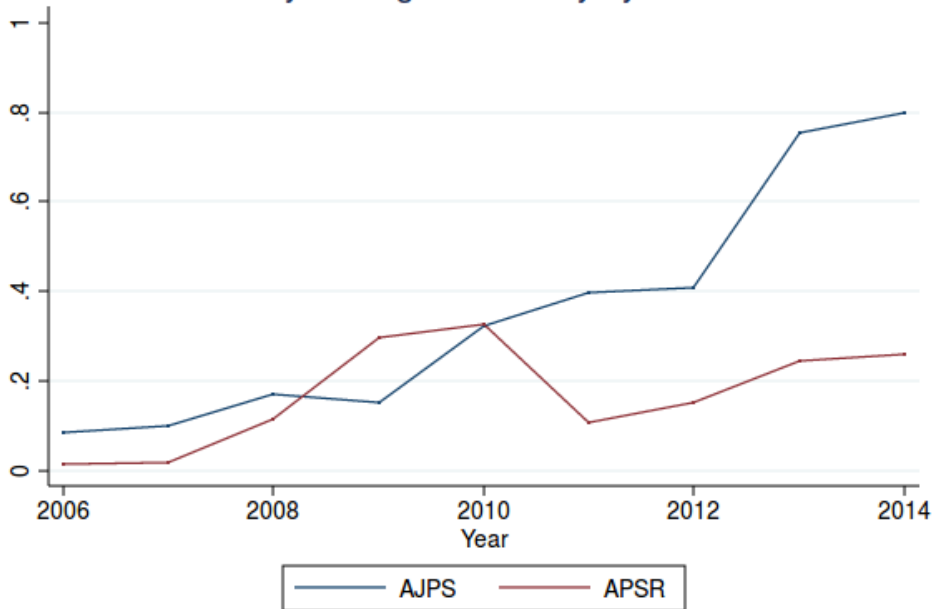
## Density of Citations



# Yearly Average Citations by Journal



Yearly Average Availability by Journal





- Naive OLS results
- First stage
- 2SLS
- Exclusion Restriction

## Naive OLS Regression

VARIABLES	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Data and Code Available	-2.039 (3.129)	13.37*** (3.271)	6.772** (3.271)	11.35*** (3.627)
AJPS		-1.533 (2.726)	-7.336*** (2.730)	-9.019*** (3.295)
Months since Pub'd		1.906* (1.120)	1.631 (1.087)	2.311* (1.324)
Months since Pub'd <sup>2</sup>		-0.0143 (0.0145)	-0.0127 (0.0141)	-0.0202 (0.0172)
Months since Pub'd <sup>3</sup>		4.51e-05 (5.82e-05)	4.66e-05 (5.65e-05)	7.83e-05 (6.94e-05)
No Data in Article			-26.80*** (3.306)	
Constant	34.52*** (1.646)	-50.90* (26.58)	-30.53 (25.89)	-54.37* (31.55)
Observations	979	979	974	745
R-squared	0.000	0.131	0.188	0.188
Sample	All	All	All	Data-Only

## Naive OLS Regression

VARIABLES	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Data and Code Available	-2.039 (3.129)	13.37*** (3.271)	6.772** (3.271)	11.35*** (3.627)
AJPS		-1.533 (2.726)	-7.336*** (2.730)	-9.019*** (3.295)
Months since Pub'd		1.906* (1.120)	1.631 (1.087)	2.311* (1.324)
No Data in Article			-26.80*** (3.306)	
Observations	979	979	974	745
R-squared	0.000	0.131	0.188	0.188
Sample	All	All	All	Data-Only

Standard errors in parentheses

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

## Naive OLS Regression

VARIABLES	(1) Ln(Cites+1)	(2) Ln(Cites+1)	(3) Ln(Cites+1)	(4) Ln(Cites+1)
Data and Code Available	0.0143 (0.0829)	0.445*** (0.0841)	0.189** (0.0799)	0.277*** (0.0782)
AJPS		-0.00263 (0.0701)	-0.229*** (0.0667)	-0.251*** (0.0710)
Months since Pub'd		0.0797*** (0.0288)	0.0714*** (0.0265)	0.0894*** (0.0285)
No Data in Article			-1.015*** (0.0807)	
Observations	979	979	974	745
R-squared	0.000	0.180	0.298	0.274
Sample	All	All	All	Data-Only

Standard errors in parentheses

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

VARIABLES	(1) First Stage	(2) First Stage	(3) First Stage
AJPS Post-2010 with Data		0.203*** (0.0695)	
AJPS Post-2012 with Data		0.287*** (0.0797)	
AJPS	0.0259 (0.0343)	-0.0121 (0.0323)	-0.0300 (0.0464)
Post-Oct 2010	-0.225*** (0.0788)	-0.147 (0.0905)	-0.245** (0.0981)
Post-July 2012	-0.0955 (0.0787)	-0.303*** (0.0985)	-0.107 (0.0970)
Post-2010 with Data		-0.0478 (0.0899)	
Post-2012 with Data		0.234** (0.102)	
Months since Pub'd	-0.00568 (0.0131)	-0.0132 (0.0125)	-0.0201 (0.0162)
No Data in Article		-0.165*** (0.0389)	
AJPS post-2010 Policy	0.205*** (0.0664)		0.219*** (0.0839)
AJPS post-2012 Policy	0.268*** (0.0738)		0.289*** (0.0905)
Observations	988	983	740
R-squared	0.257	0.336	0.261
Sample	All	IV=Data-Only	Data-Only
F Stat	33.27	33.61	23.58

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

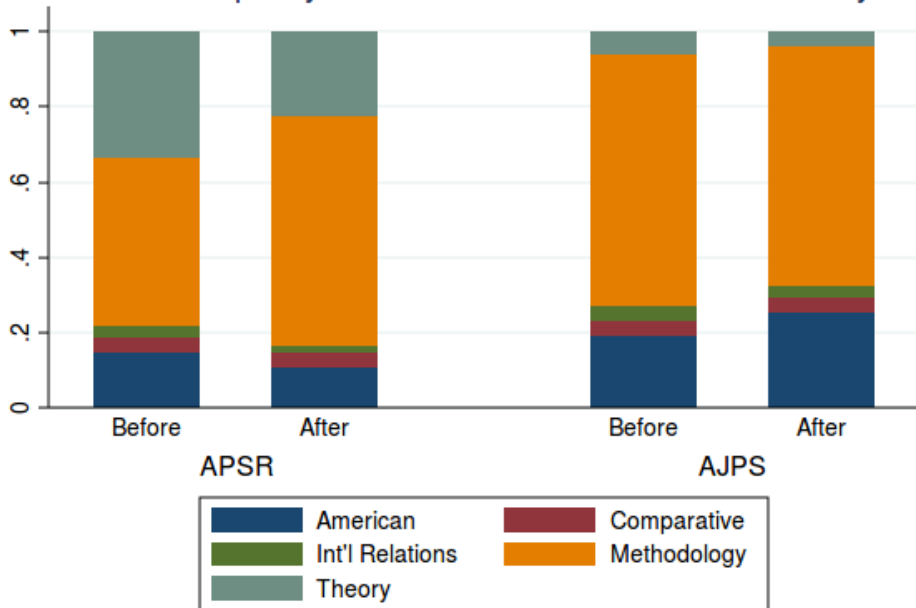
## 2SLS Regression

VARIABLES	(1) 2SLS	(2) 2SLS	(3) 2SLS
Data and Code Available	0 (0)	0 (0)	-3.472 (13.10)
AJPS	0.225 (2.605)	-7.288*** (2.637)	-7.013* (4.036)
Post-Oct 2010	-21.38*** (7.381)	1.211 (9.578)	-22.18** (8.780)
Post-July 2012	-12.70** (5.814)	-2.670 (9.417)	-12.35* (7.177)
Post-2010 with Data		-21.84*** (8.317)	
Post-2012 with Data		-8.412 (8.867)	
Months since Pub'd	2.204*** (0.716)	2.282*** (0.669)	2.637*** (0.920)
No Data in Article		-39.52*** (4.124)	
Observations	979	974	745
R-squared	0.123	0.206	0.176
Sample	All	IV=Data-Only	Data-Only

2SLS Regression of  $\ln(\text{citations}+1)$

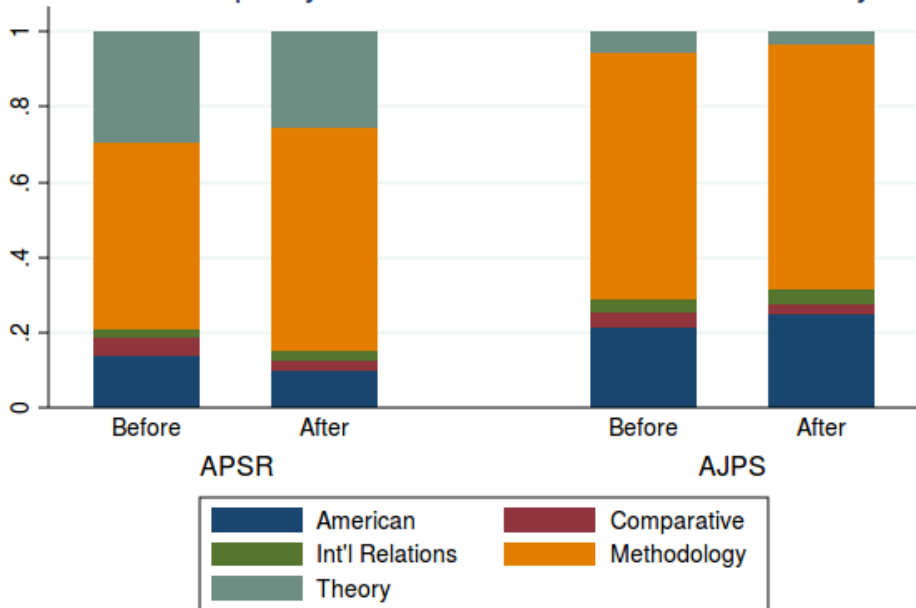
VARIABLES	(1) 2SLS-Log	(2) 2SLS-Log	(3) 2SLS-Log
Data and Code Available	0 (0)	0 (0)	0.131 (0.281)
AJPS	0.0780 (0.0676)	-0.208*** (0.0650)	-0.229*** (0.0866)
Post-Oct 2010	-0.179 (0.192)	0.368 (0.236)	-0.186 (0.188)
Post-July 2012	-0.0754 (0.151)	-0.00629 (0.232)	-0.0995 (0.154)
Post-2010 with Data		-0.471** (0.205)	
Post-2012 with Data		0.0403 (0.219)	
Months since Pub'd	0.0859*** (0.0186)	0.0890*** (0.0165)	0.100*** (0.0197)
No Data in Article		-1.250*** (0.102)	
Observations	979	974	745
R-squared	0.157	0.300	0.272
Sample	All	IV=Data-Only	Data-Only

Article Topic by Journal Before and After 2010 Policy

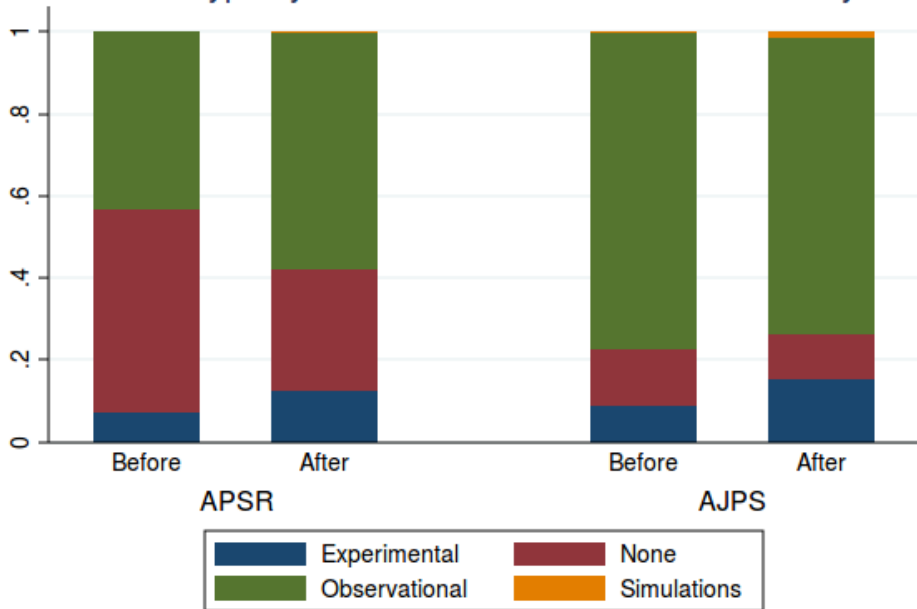




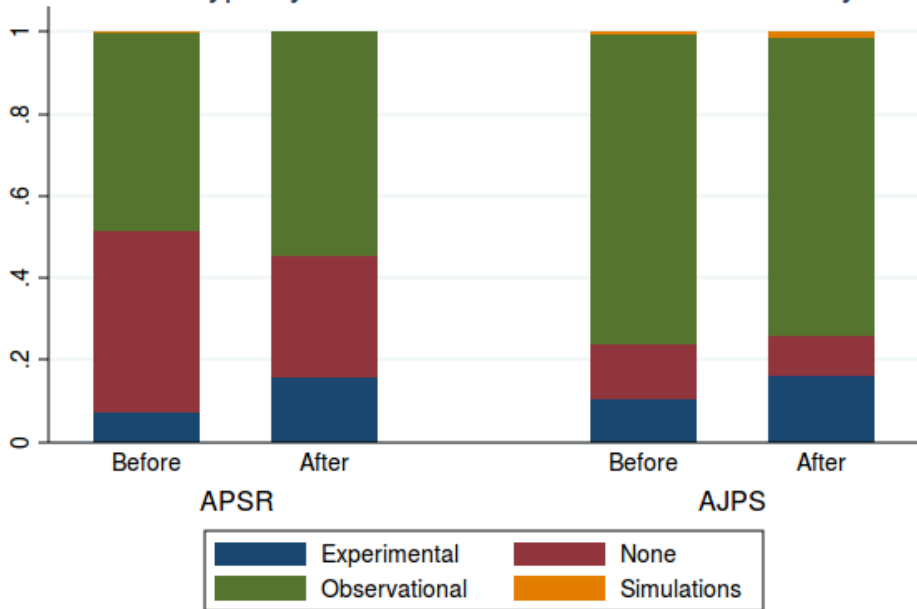
Article Topic by Journal Before and After 2012 Policy



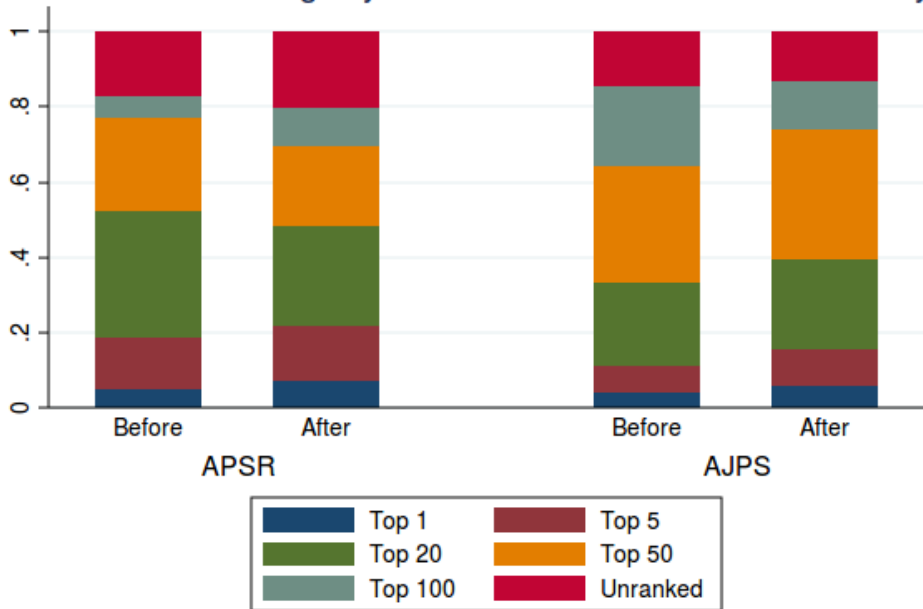
Data Type by Journal Before and After 2010 Policy



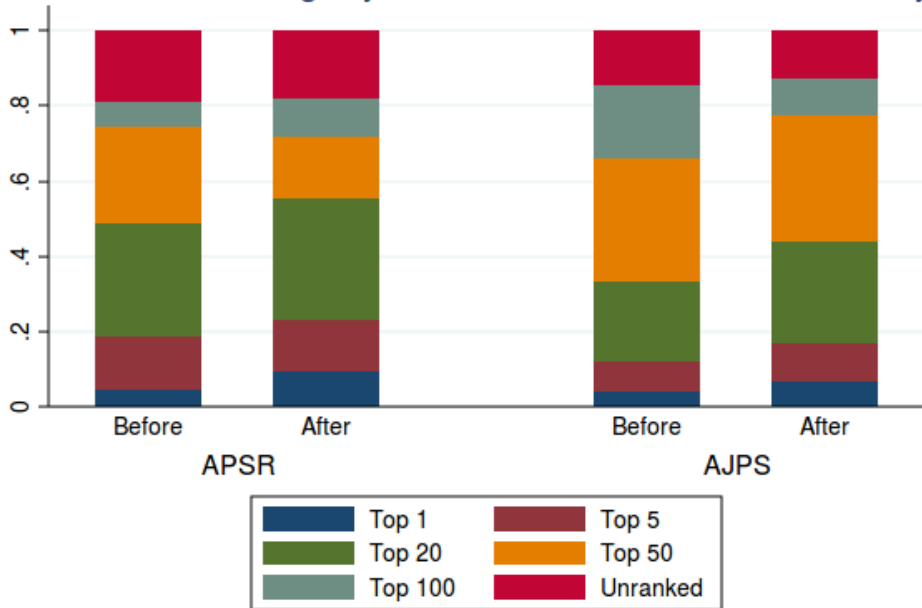
Data Type by Journal Before and After 2012 Policy



Institution Rankings by Journal Before and After 2010 Policy



Institution Rankings by Journal Before and After 2012 Policy



VARIABLES	Exclusion Restriction					
	(1) American	(2) Methodology	(3) Experimental	(4) Observational	(5) Top 5	(6) Top 20
AJPS post-2010 Policy	0.134 (0.0827)	-0.163* (0.0928)	0.102 (0.0709)	-0.0993 (0.0725)	9.98e-05 (0.0595)	0.160* (0.0887)
AJPS post-2012 Policy	-0.00687 (0.0891)	0.00991 (0.100)	-0.105 (0.0764)	0.0919 (0.0782)	0.00406 (0.0645)	0.0139 (0.0961)
AJPS	0.0287 (0.0457)	0.0254 (0.0513)	-0.0419 (0.0392)	0.0376 (0.0401)	-0.0594* (0.0330)	-0.201*** (0.0492)
Post-Oct 2010	-0.0168 (0.0966)	0.00684 (0.109)	-0.000582 (0.0828)	-0.000418 (0.0847)	0.0267 (0.0694)	-0.0949 (0.103)
Post-July 2012	-0.0513 (0.0956)	0.0786 (0.107)	0.163** (0.0820)	-0.153* (0.0838)	0.0285 (0.0690)	0.0744 (0.103)
Months since Pub'd	-0.0251 (0.0160)	0.0331* (0.0179)	-0.00167 (0.0137)	-0.00439 (0.0140)	0.00289 (0.0115)	-0.00648 (0.0172)
Observations	740	740	740	740	733	733
R-squared	0.022	0.017	0.015	0.016	0.011	0.028
Sample	Data-Only	Data-Only	Data-Only	Data-Only	Data-Only	Data-Only

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

# Preliminary Conclusions

- Top political science papers with public data are cited more.
- Some suggestive, not strong, evidence of causality.
- Journal policy does not appear to have changed submissions.
  - IV identification strategy OK.

# Future

- Data quality checks.
- Economics: AER & QJE 2001-2009