BERKELEY INITIATIVE FOR TRANSPARENCY
IN THE SOCIAL SCIENCES

Data Sharing and Citations

Causal Evidence

Garret Christensen[1] Allan Dafoe[2] Edward Miguel[3]

[1]Berkeley Institute for Data Science, UC Berkeley

[2]Department of Political Science, Yale University

[3]Department of Economics, UC Berkeley

December 2017 BITSS Annual Meeting

PRELIMINARY–Please do not cite.

Thanks to David Birke, Mu Yang Shin, Don Sun, Manana Hakobyan, Terri Cruz, Maxim Guzman, Baiyue Cao, Evey Huang, Rachel Kim, Ravina Pattni, Kevin Khuu

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.

Google joshua angrist

Scholar About 11,500 results (0.09 sec)

Articles
Case law
My library

Any time
Since 2017
Since 2016
Since 2013
Custom range...

Sort by relevance
Sort by date

include patents
include citations

Create alert

[BOOK] Mostly harmless econometrics: An empiricist's companion
JD Angrist, JS Pischke - 2008 - books.google.com
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

Identification of causal effects using instrumental variables [PDF] zmjones.com
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

Identification and estimation of local average treatment effects [PDF] baylor.edu
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

Does compulsory school attendance affect schooling and earnings? [PDF] princeton.edu
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

Instrumental variables and the search for identification: From supply and demand to natural experiments [PDF] mit.edu
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

Using Maimonides' rule to estimate the effect of class size on scholastic achievement [PDF] ens.fr

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?”

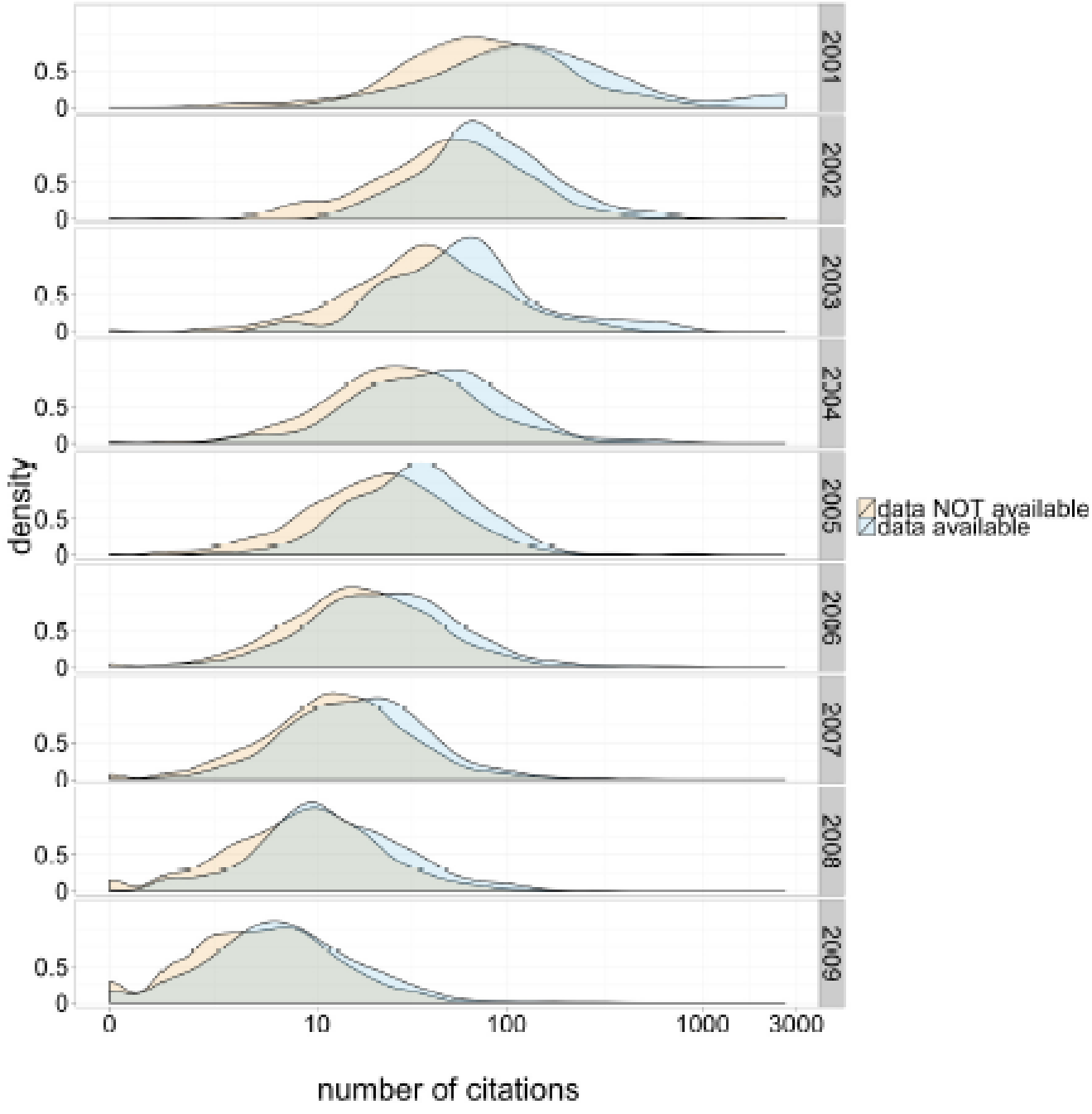
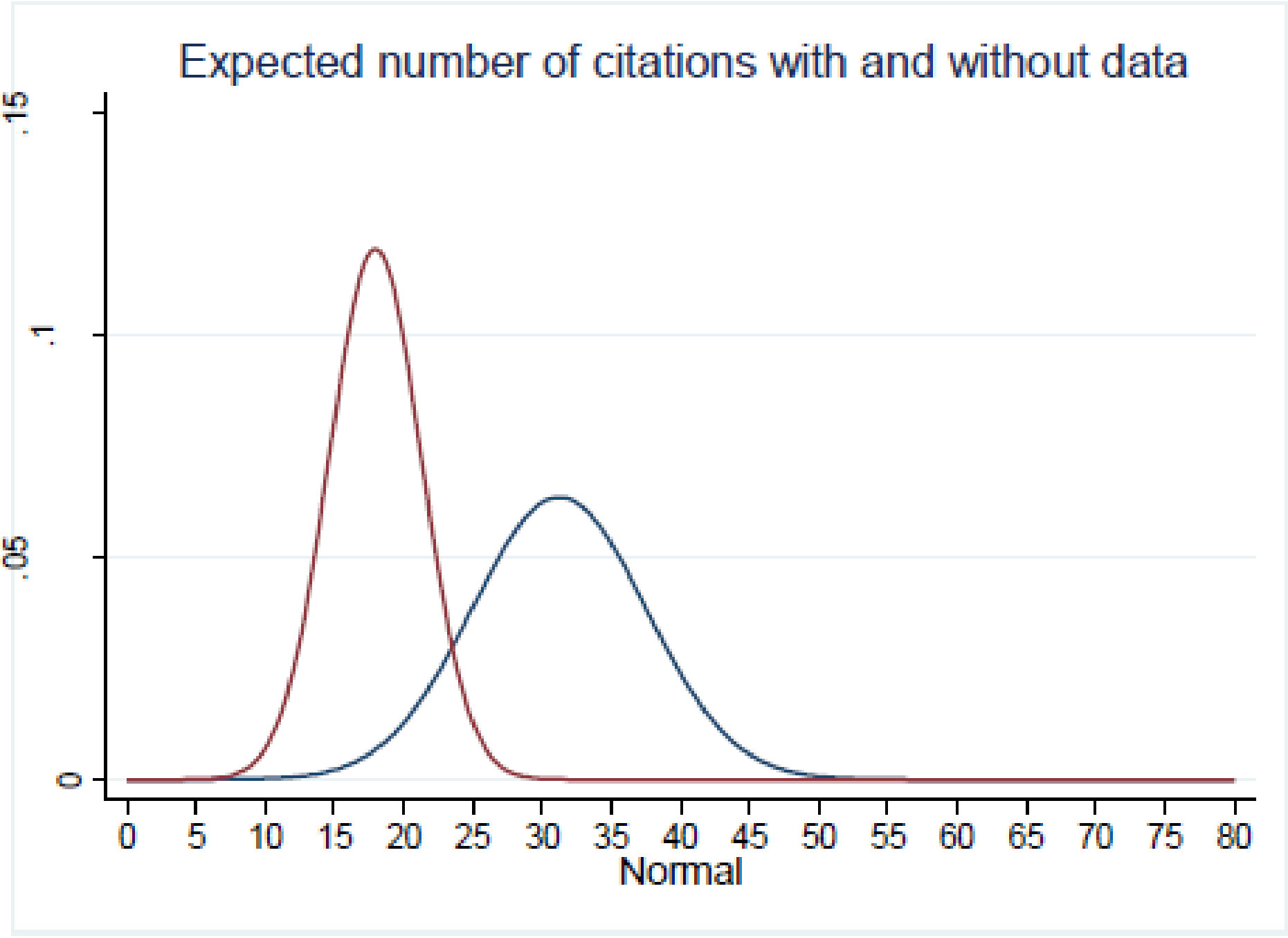


Figure 1 **Citation density for papers with and without publicly available microarray data, by year of study publication.**

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.

$$corr(Z, u) = 0$$

$$corr(Z, X) \neq 0$$

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, $corr(Z, X) \neq 0; F > 10$)
- Exclusion Restriction. ($corr(Z, u) = 0$; 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.

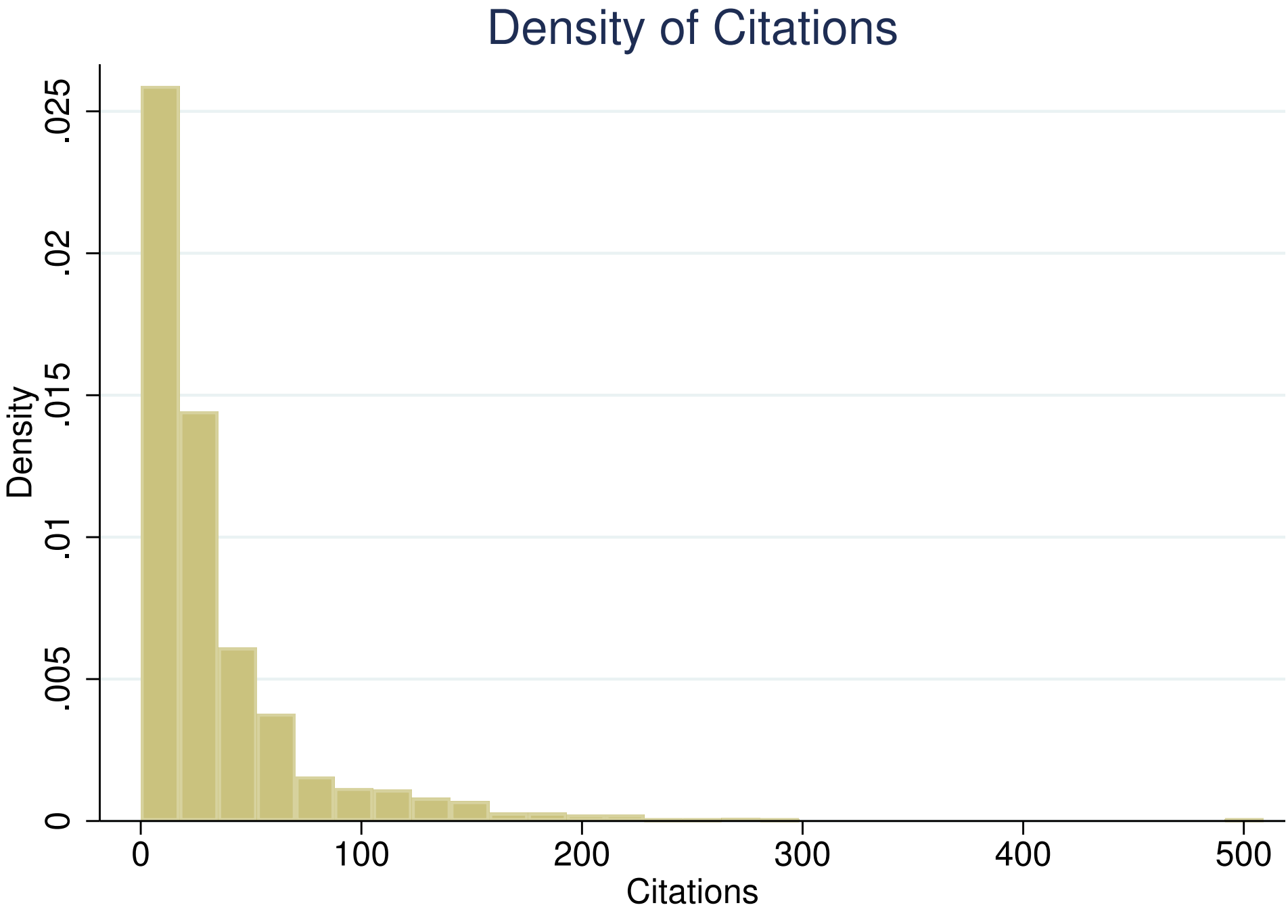
Our Estimation Strategy

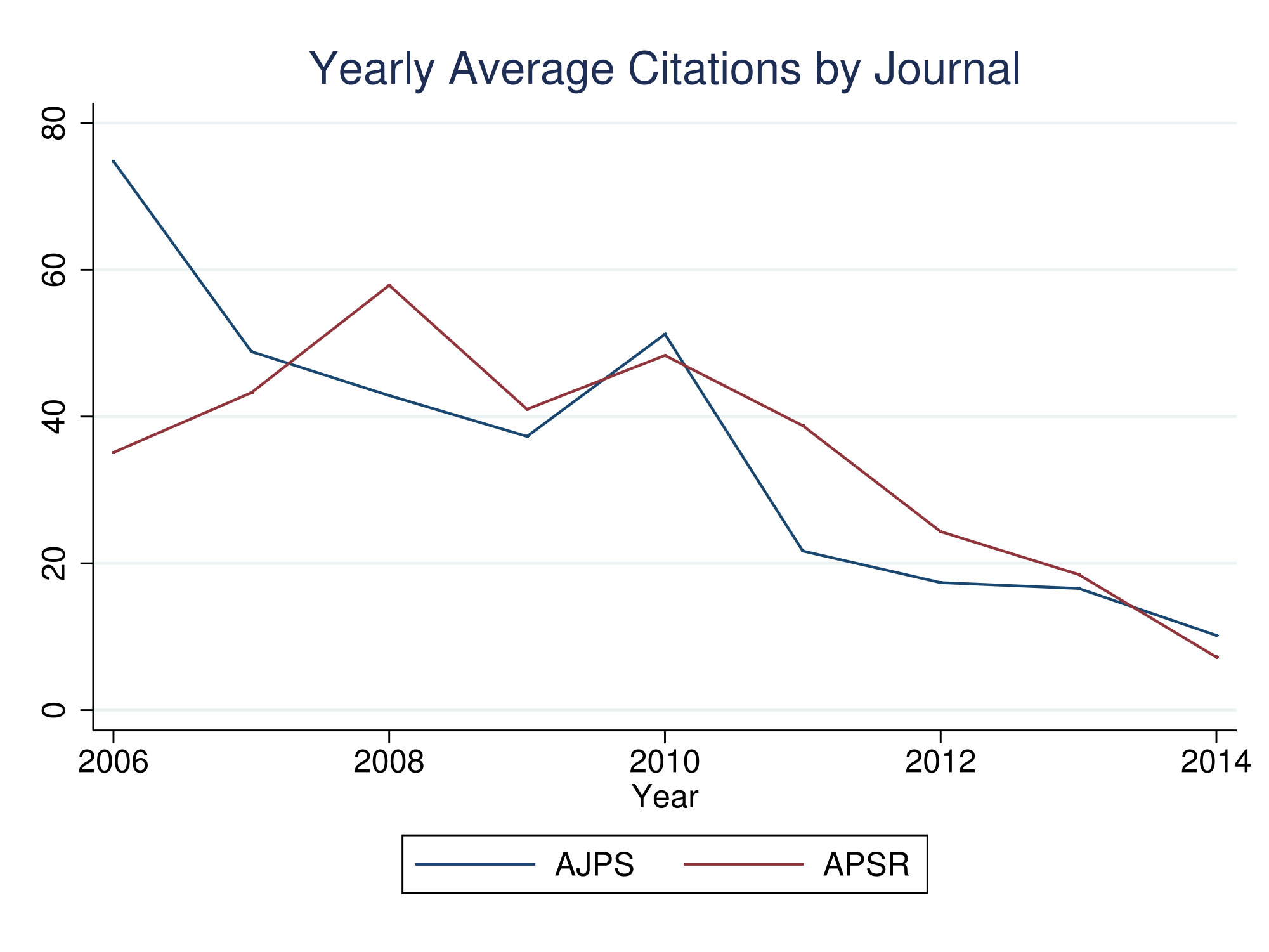
$$\begin{aligned} availability_i = \alpha_1 + \beta_1 AJPS_i + \beta_2 Post2010_i + \beta_3 Post2012_i \quad (1) \\ + \beta_4 AJPS * Post2010_i + \beta_5 AJPS * Post2012_i \\ + g_1(Time) + h_1(Year) + \nu_i \end{aligned}$$

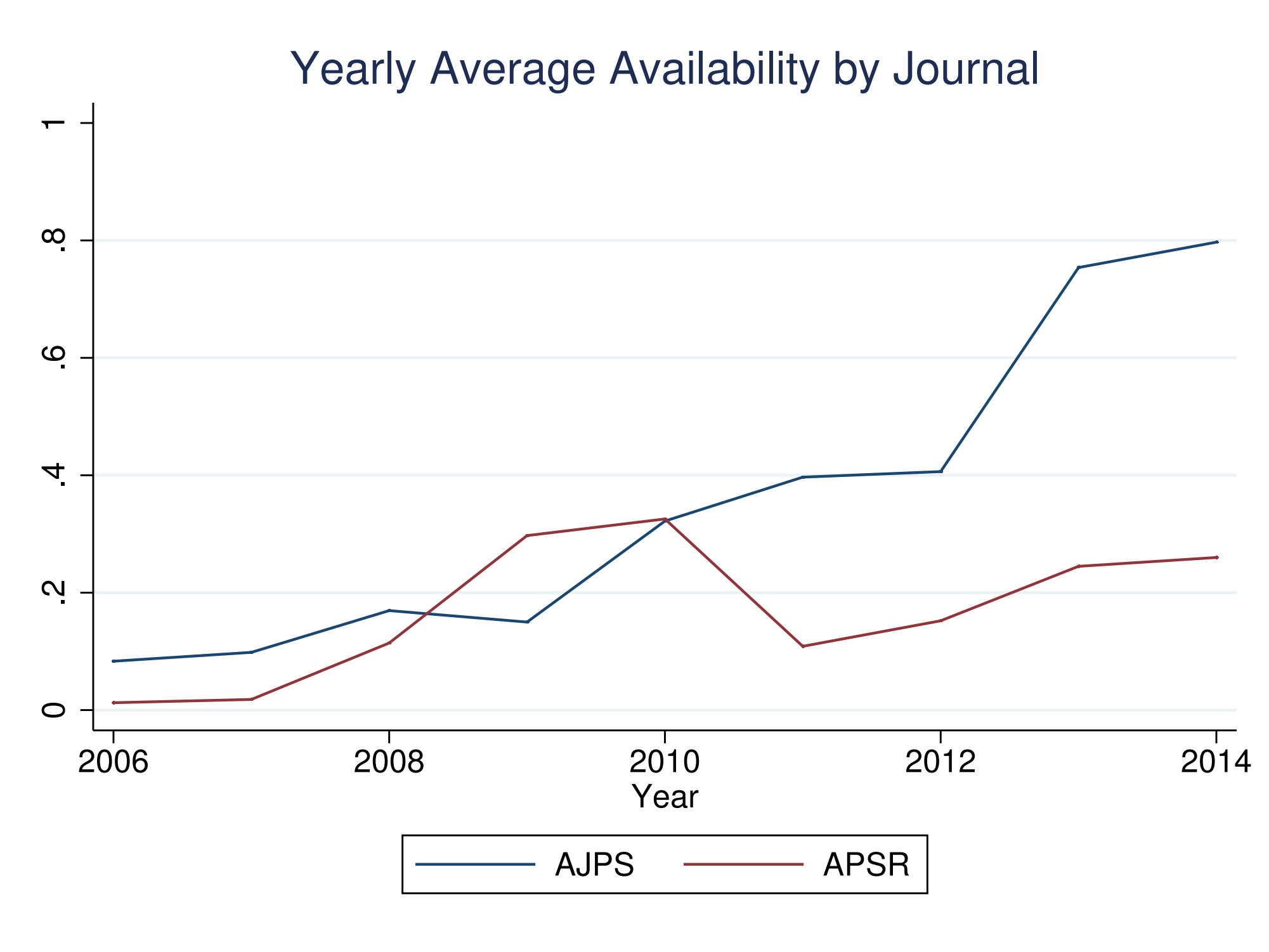
$$\begin{aligned} citations_i = \alpha_2 + \eta_1 AJPS_i + \eta_2 Post2010_i + \eta_3 Post2012_i \quad (2) \\ + \eta_4 \hat{availability}_i + g_2(Time) + h_2(Year) + u_i \end{aligned}$$

Summary Statistics

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


Density of Citations
Density
0
.005
.01
.015
.02
.025
0
100
200
300
400
500
Citations


Yearly Average Citations by Journal
0
20
40
60
80
2006
2008
2010
2012
2014
Year
AJPS
APSR


Yearly Average Availability by Journal
0
.2
.4
.6
.8
1
2006
2008
2010
2012
2014
Year
AJPS
APSR

- 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	13.371***	6.772**	11.354***
	(3.129)	(3.271)	(3.271)	(3.627)
AJPS		-1.533	-7.336***	-9.019***
		(2.726)	(2.730)	(3.295)
Months since Pub'd		1.906*	1.631	2.311*
		(1.120)	(1.087)	(1.324)
Months since Pub'd^2		-0.014	-0.013	-0.020
		(0.014)	(0.014)	(0.017)
Months since Pub'd^3		0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)
No Data in Article			-26.799***	
			(3.306)	
Constant	34.523***	-50.900*	-30.534	-54.373*
	(1.646)	(26.577)	(25.890)	(31.554)
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	13.371***	6.772**	11.354***
	(3.129)	(3.271)	(3.271)	(3.627)
AJPS		-1.533	-7.336***	-9.019***
		(2.726)	(2.730)	(3.295)
Months since Pub'd		1.906*	1.631	2.311*
		(1.120)	(1.087)	(1.324)
No Data in Article			-26.799***	
			(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	0.014	0.445***	0.189**	0.277***
	(0.083)	(0.084)	(0.080)	(0.078)
AJPS		-0.003	-0.229***	-0.251***
		(0.070)	(0.067)	(0.071)
Months since Pub'd		0.080***	0.071***	0.089***
		(0.029)	(0.027)	(0.029)
No Data in Article			-1.015***	
			(0.081)	
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 Policy	0.205***		0.219***
	(0.066)		(0.084)
AJPS post-2012 Policy	0.268***		0.289***
	(0.074)		(0.091)
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

Controls cut for space

2SLS Regression

VARIABLES	(1) 2SLS	(2) 2SLS	(3) 2SLS
Data and Code Available	-20.012	1.701	3.861
	(13.553)	(13.004)	(14.617)
AJPS	4.601	-7.338**	-8.098**
	(3.862)	(3.220)	(4.119)
Post-Oct 2010	-24.914***	0.579	-22.280**
	(7.879)	(9.862)	(8.705)
Post-July 2012	-6.663	0.869	-9.132
	(7.204)	(11.184)	(7.756)
Post-2010 with Data		-21.802***	
		(8.340)	
Post-2012 with Data		-9.284	
		(10.799)	
Months since Pub'd	3.669**	3.267**	4.078**
	(1.449)	(1.335)	(1.658)
No Data in Article		-39.114***	
		(4.320)	
Observations	979	974	745
R-squared	0.048	0.208	0.191
Sample	All	IV=Data-Only	Data-Only

2SLS Regression of ln(citations+1)

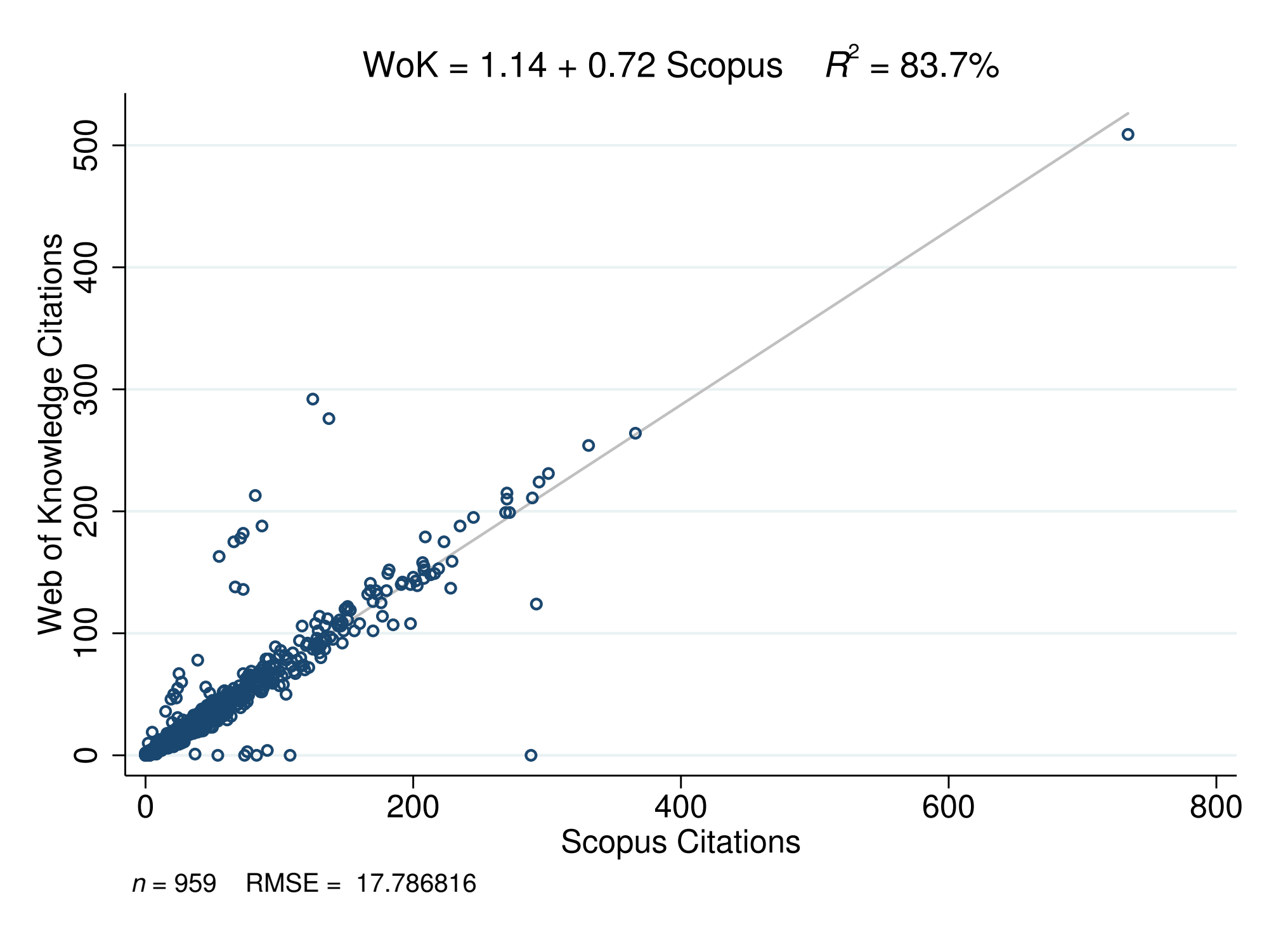
VARIABLES	(1) 2SLS-Log	(2) 2SLS-Log	(3) 2SLS-Log
Data and Code Available	-0.495	0.025	0.125
	(0.353)	(0.321)	(0.316)
AJPS	0.180*	-0.215***	-0.228**
	(0.101)	(0.079)	(0.089)
Post-Oct 2010	-0.242	0.392	-0.185
	(0.205)	(0.243)	(0.188)
Post-July 2012	-0.037	-0.051	-0.102
	(0.188)	(0.276)	(0.168)
Post-2010 with Data		-0.474**	
		(0.206)	
Post-2012 with Data		0.024	
		(0.266)	
Months since Pub'd	0.087**	0.069**	0.099***
	(0.038)	(0.033)	(0.036)
No Data in Article		-1.252***	
		(0.107)	
Observations	979	974	745
R-squared	0.076	0.302	0.272
Sample	All	IV=Data-Only	Data-Only

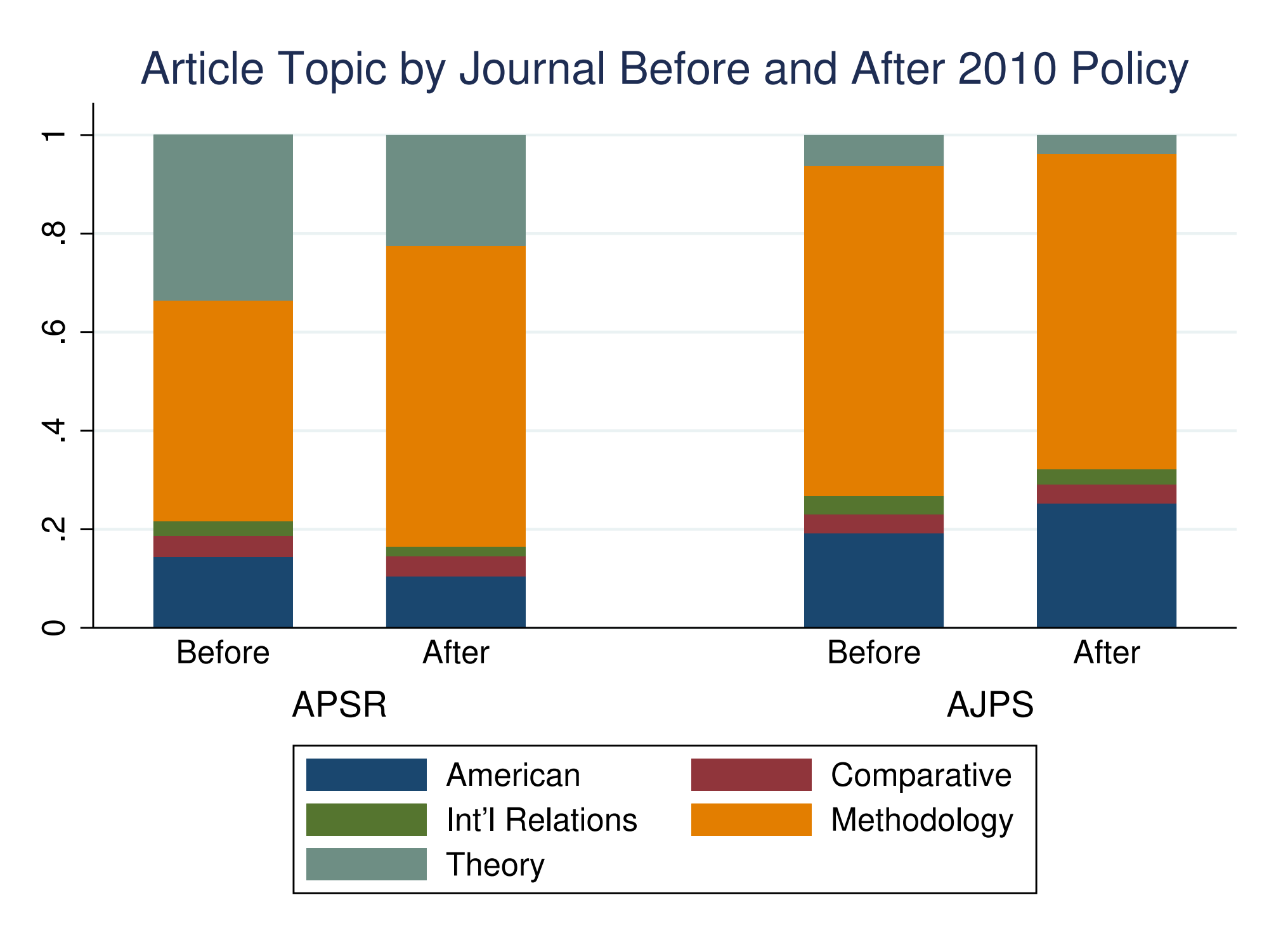
Citation Data

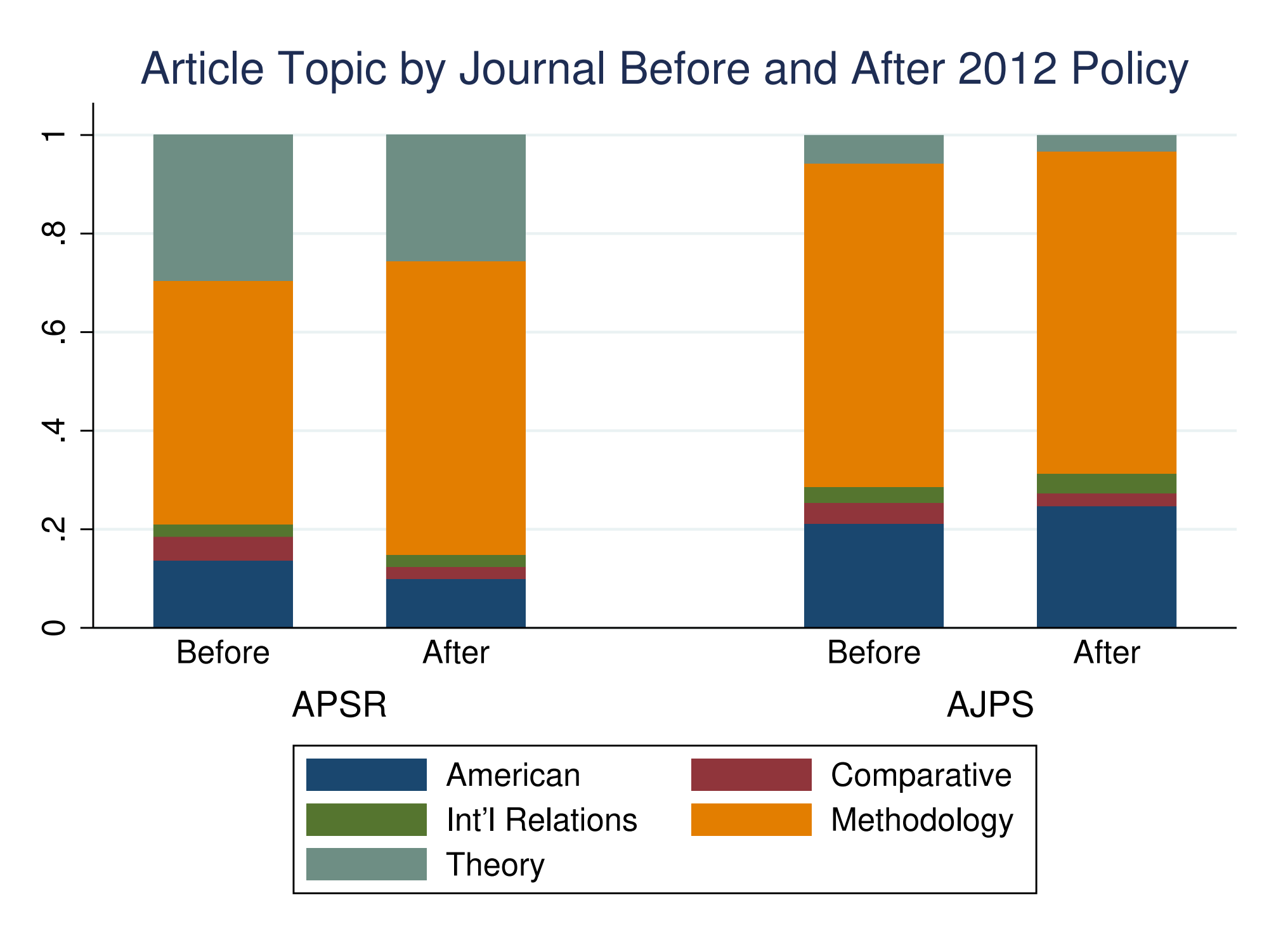
Citations are interesting data in and of themselves.

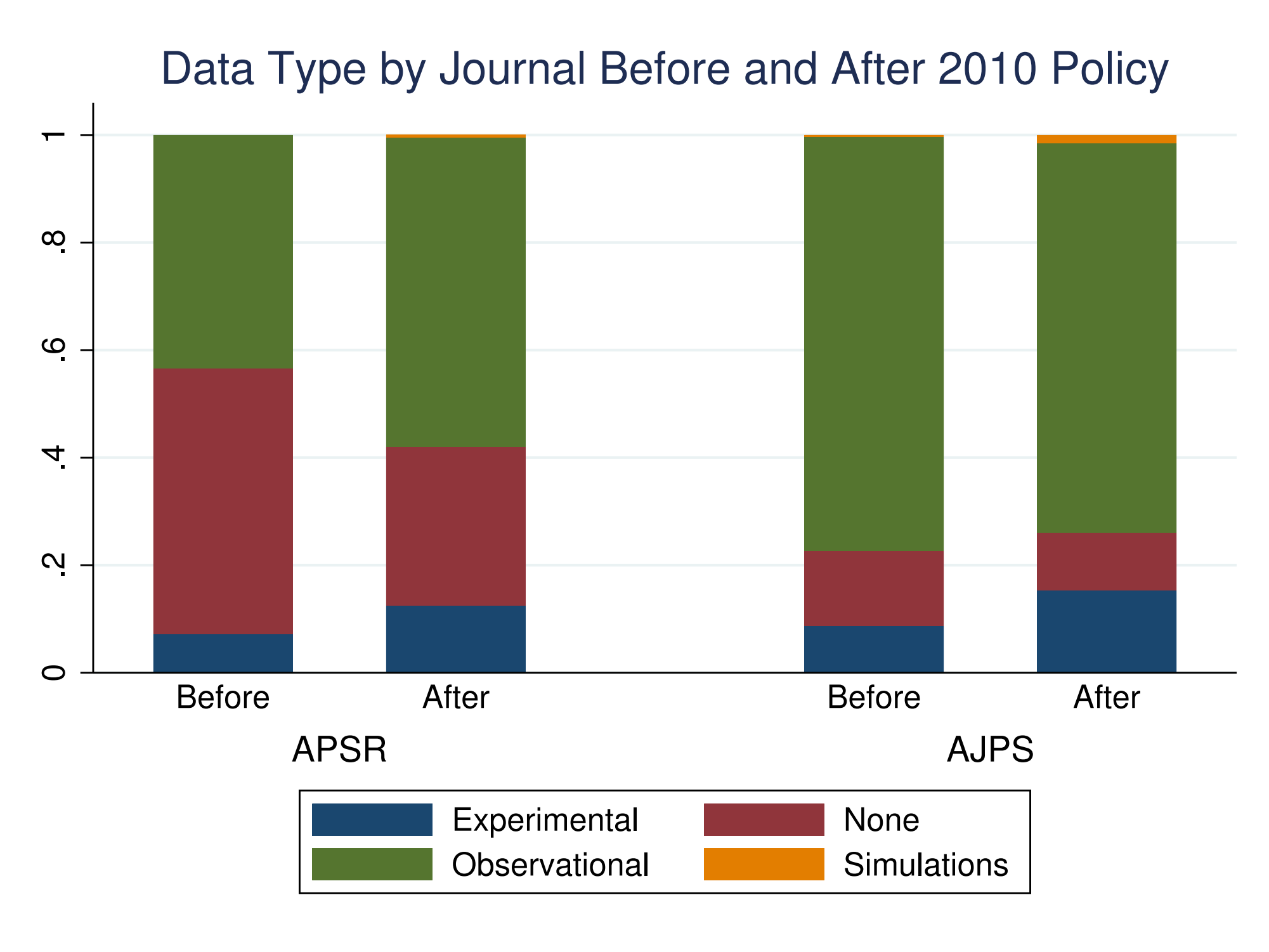
- Multiple sources
- Not open
- Initiative to change that: I4OC

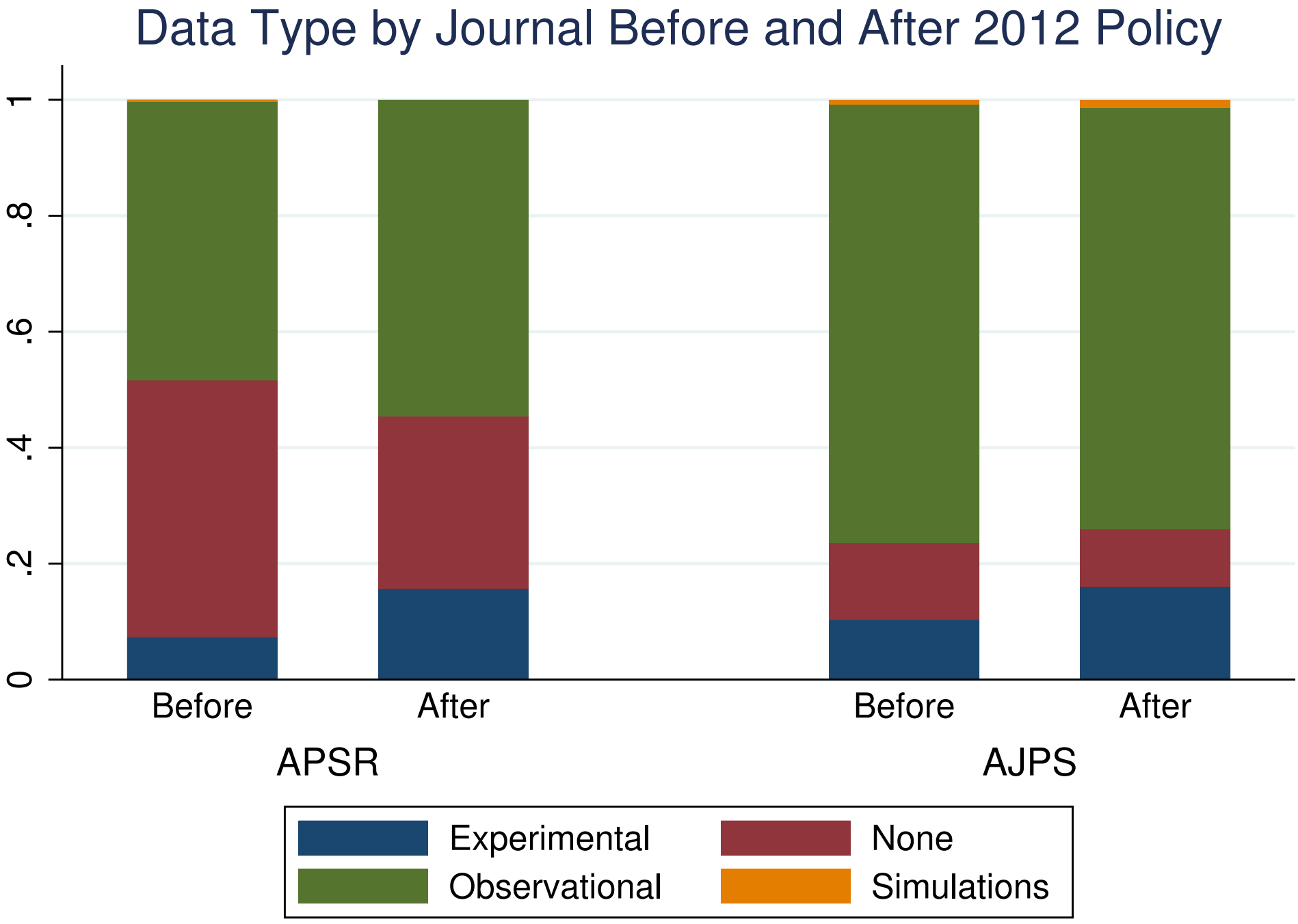
Results robust to source of citation data

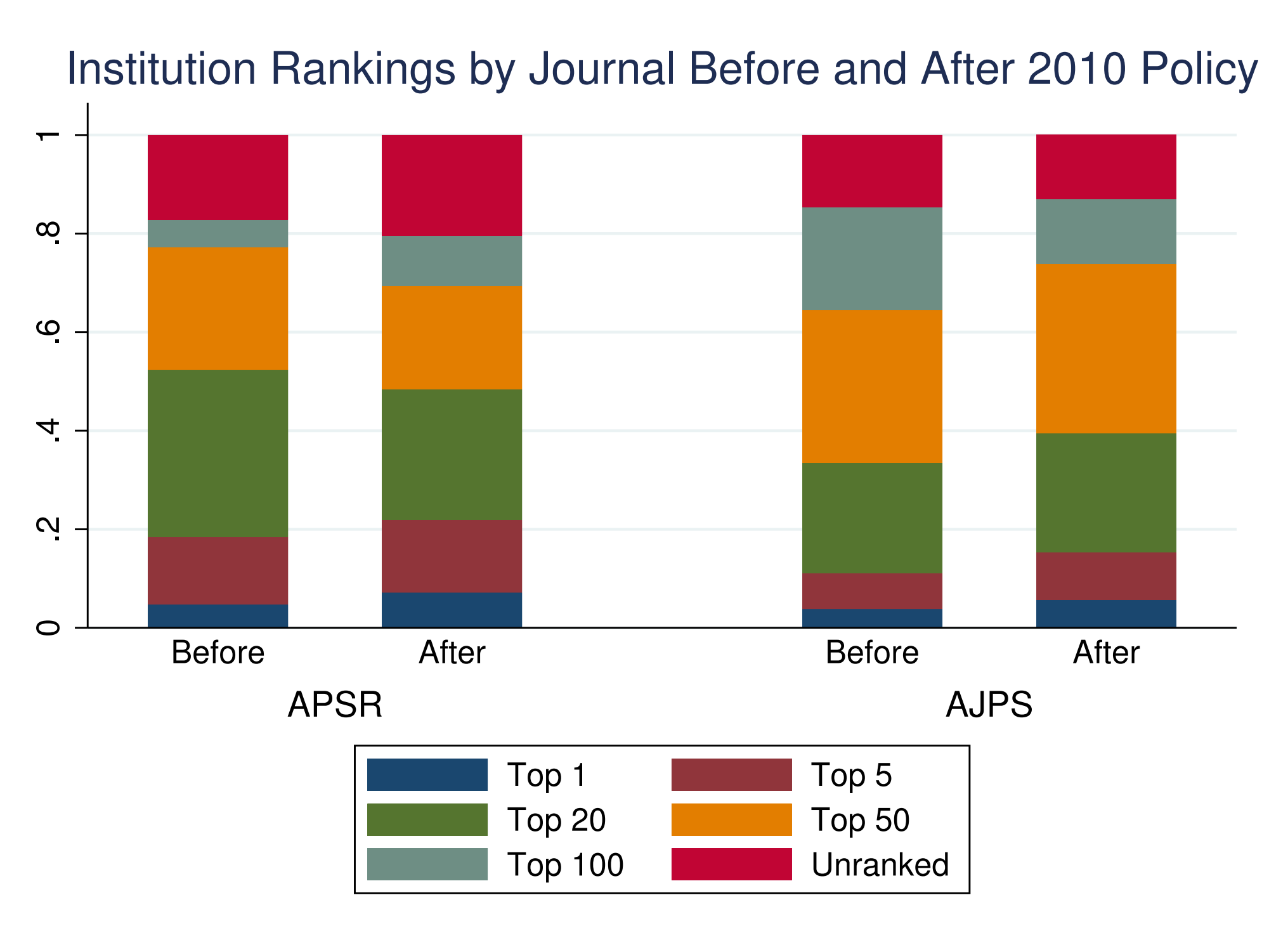

WoK = 1.14 + 0.72 Scopus R^2 = 83.7%
Web of Knowledge Citations
0
100
200
300
400
500
0
200
400
600
800
Scopus Citations
n = 959 RMSE = 17.786816

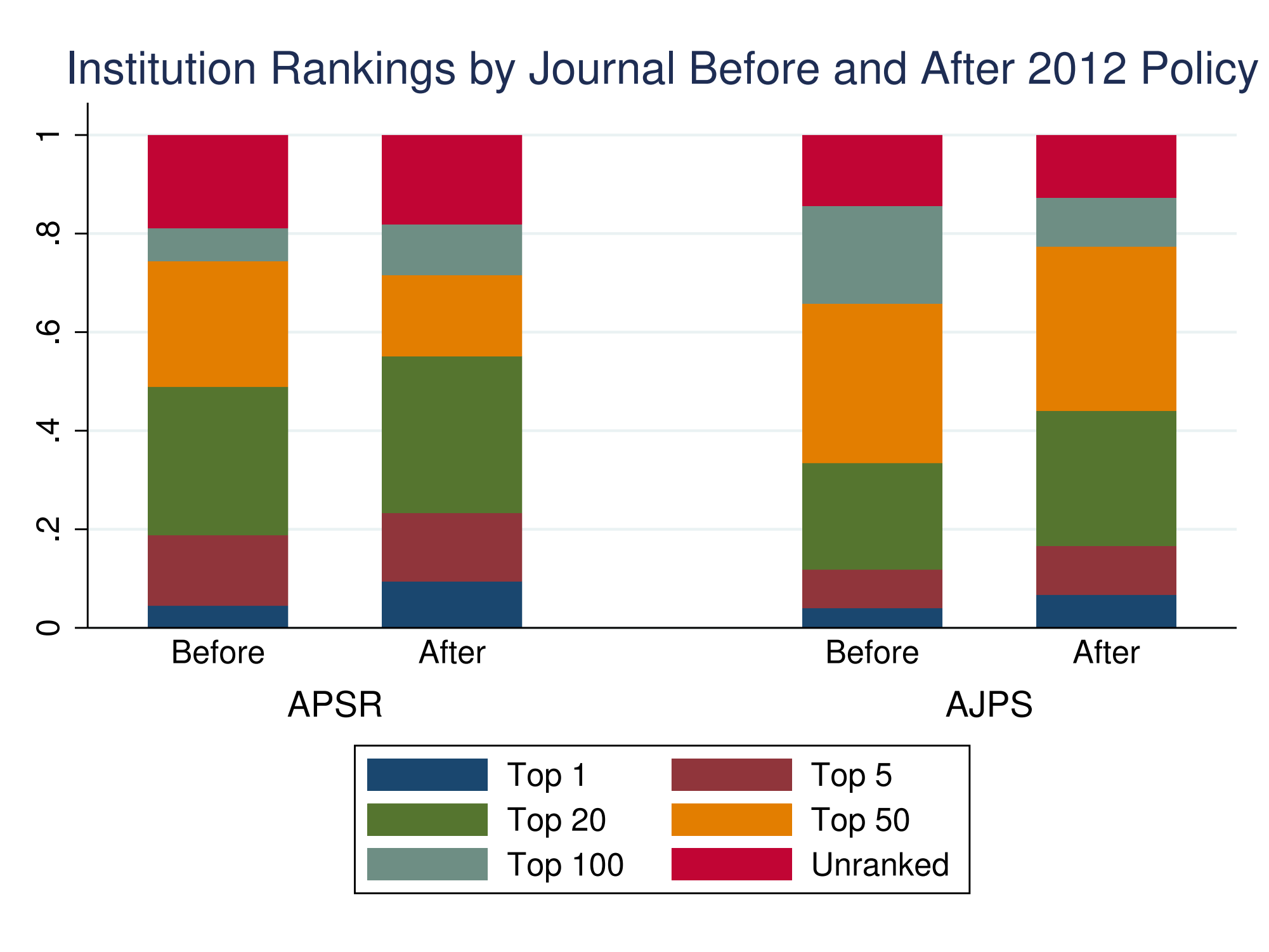

Article Topic by Journal Before and After 2010 Policy
0
.2
.4
.6
.8
1
Before
After
Before
After
APSR
AJPS
American
Comparative
Int'l Relations
Methodology
Theory


Article Topic by Journal Before and After 2012 Policy
0
.2
.4
.6
.8
1
Before
After
Before
After
APSR
AJPS
American
Comparative
Int'l Relations
Methodology
Theory


Data Type by Journal Before and After 2010 Policy
0
.2
.4
.6
.8
1
Before
After
Before
After
APSR
AJPS
Experimental
Observational
None
Simulations


Data Type by Journal Before and After 2012 Policy
0
.2
.4
.6
.8
1
Before
After
Before
After
APSR
AJPS
Experimental
Observational
None
Simulations


Institution Rankings by Journal Before and After 2010 Policy
0
.2
.4
.6
.8
1
Before
After
Before
After
APSR
AJPS
Top 1
Top 5
Top 20
Top 50
Top 100
Unranked


Institution Rankings by Journal Before and After 2012 Policy
0
.2
.4
.6
.8
1
Before
After
Before
After
APSR
AJPS
Top 1
Top 5
Top 20
Top 50
Top 100
Unranked

Exclusion Restriction

VARIABLES	(1) American	(2) Methodology	(3) Experimental	(4) Observational	(5) Top 5	(6) Top 20
AJPS post-2010 Policy	0.134	-0.163*	0.102	-0.099	0.000	0.160*
	(0.083)	(0.093)	(0.071)	(0.072)	(0.059)	(0.089)
AJPS post-2012 Policy	-0.007	0.010	-0.105	0.092	0.004	0.014
	(0.089)	(0.100)	(0.076)	(0.078)	(0.065)	(0.096)
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

Controls dropped for space

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

- Economics: AER & QJE 2001-2009
- Collaborate with Don Moore & Andrew Rose
  - Expand journals (19 with policy changes)
  - Narrow data collection