BERKELEY INITIATIVE FOR TRANSPARENCY
IN THE SOCIAL SCIENCES

Data Sharing and Citations

Causal Evidence

Garre 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

June 28, 2017 WEAI

PRELIMINARY–Please do not cite.

Outline

PRELIMINARY–Please do not cite.

Data Sharing Incentives

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

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?”

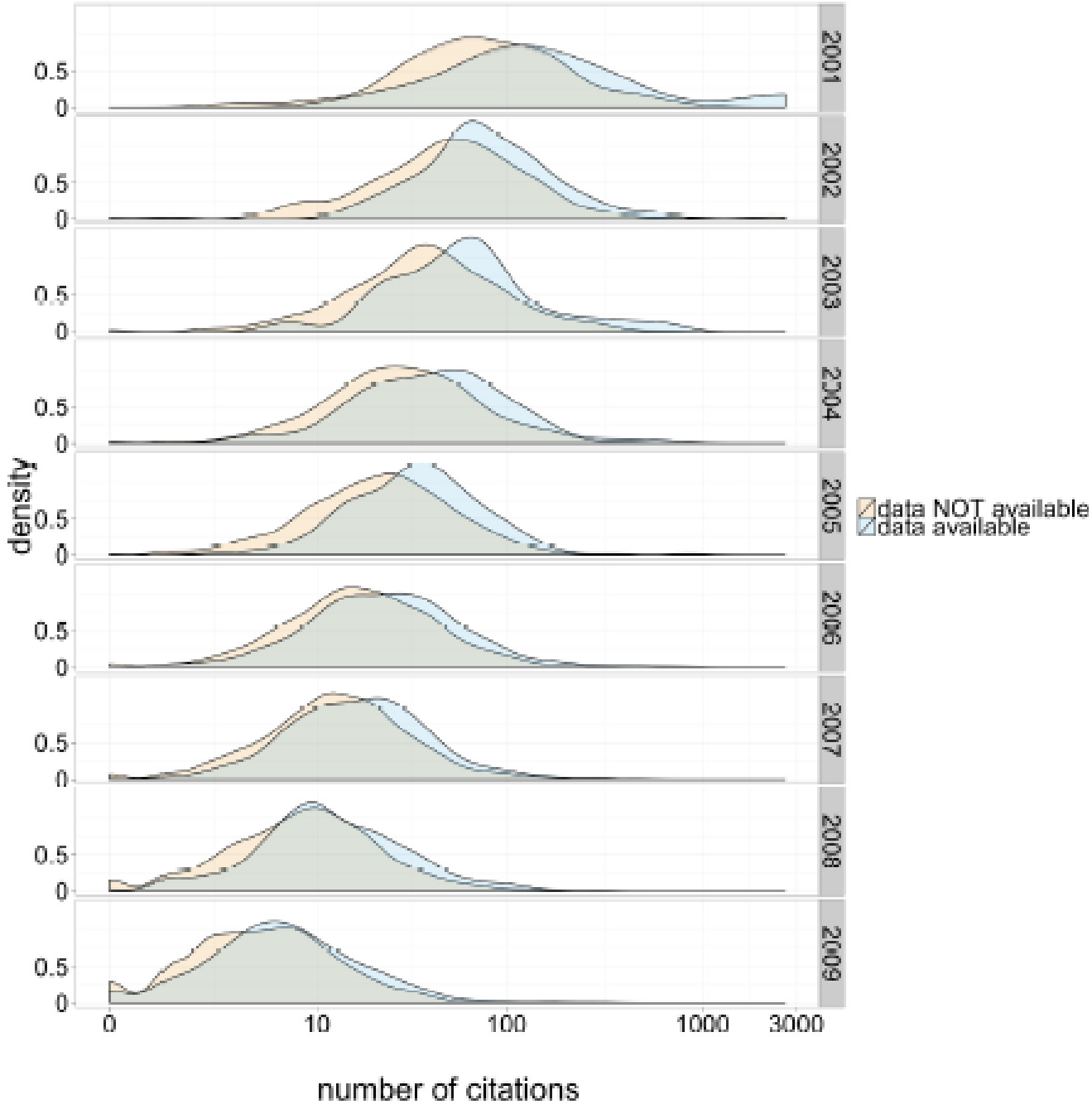
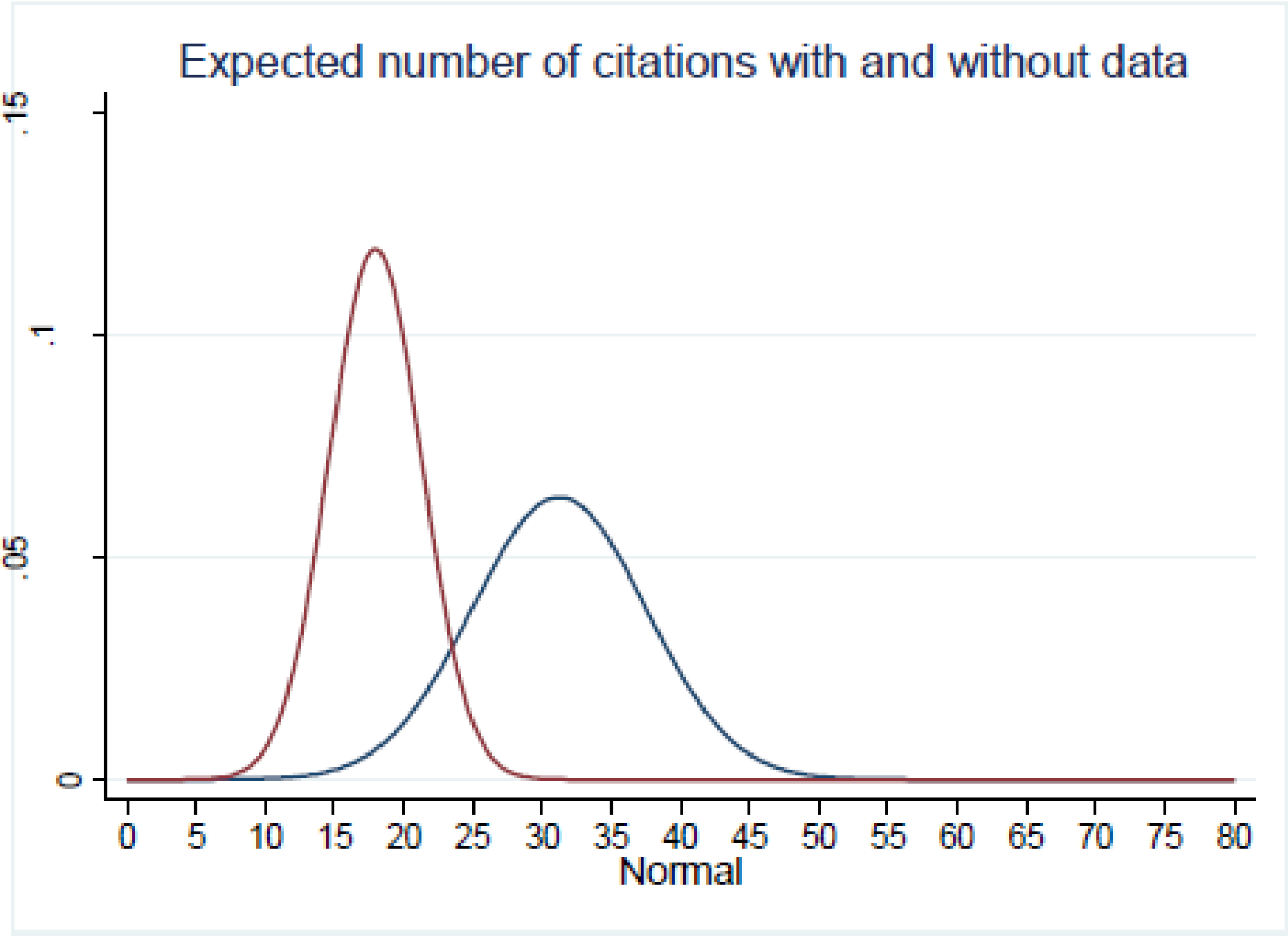


Figure 1 **Citation density for papers with and without publicly available microarray data, by year of study publication.**

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 *"If a manuscript is accepted for publication it will not be published unless the first footnote explicitly notes where the data used in the study can be obtained for purposes of replication and should note any sources that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

Pre-Analysis Plan

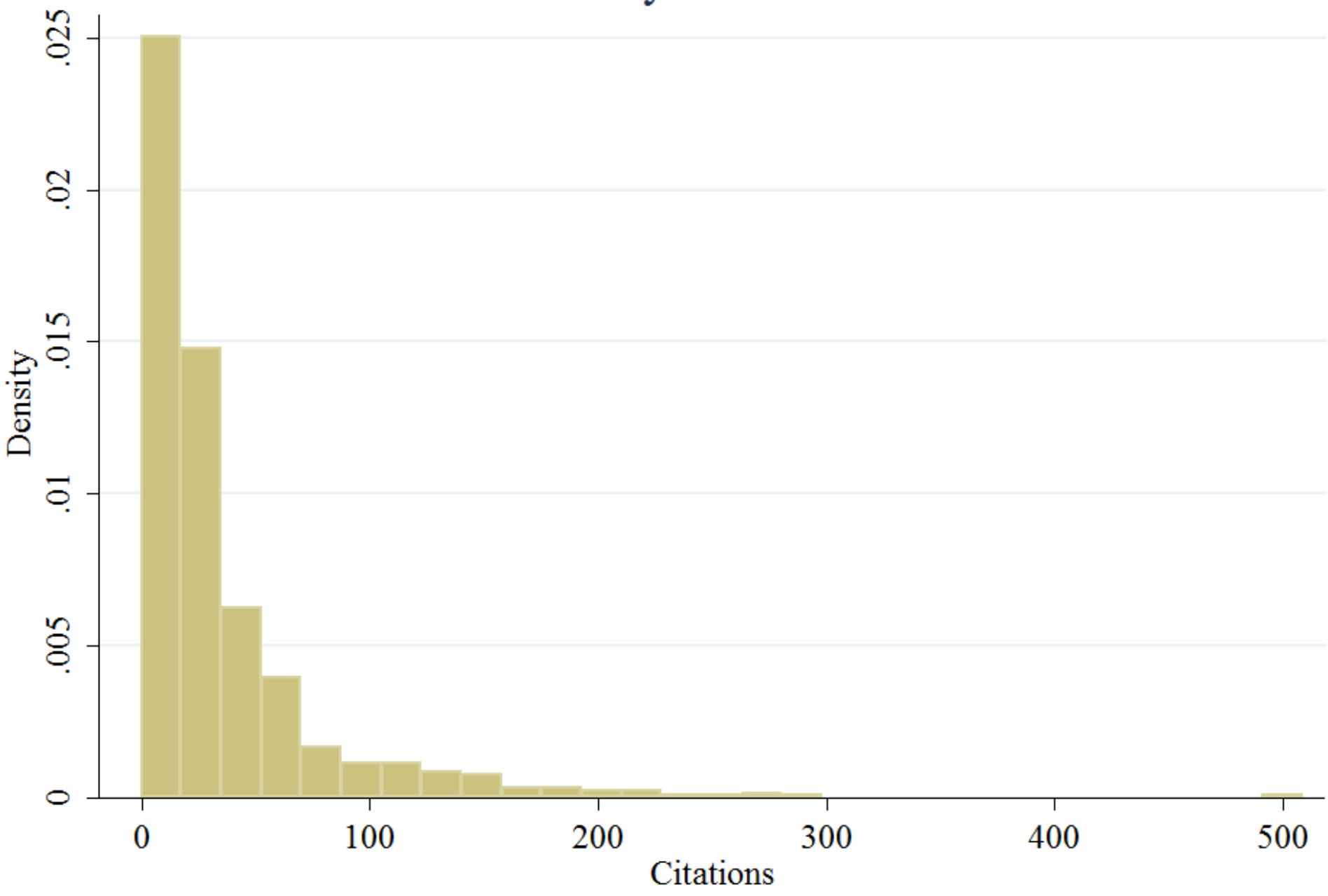
- Short pre-analysis plan before data collection.
- Available at `https://osf.io/qxpr6/`

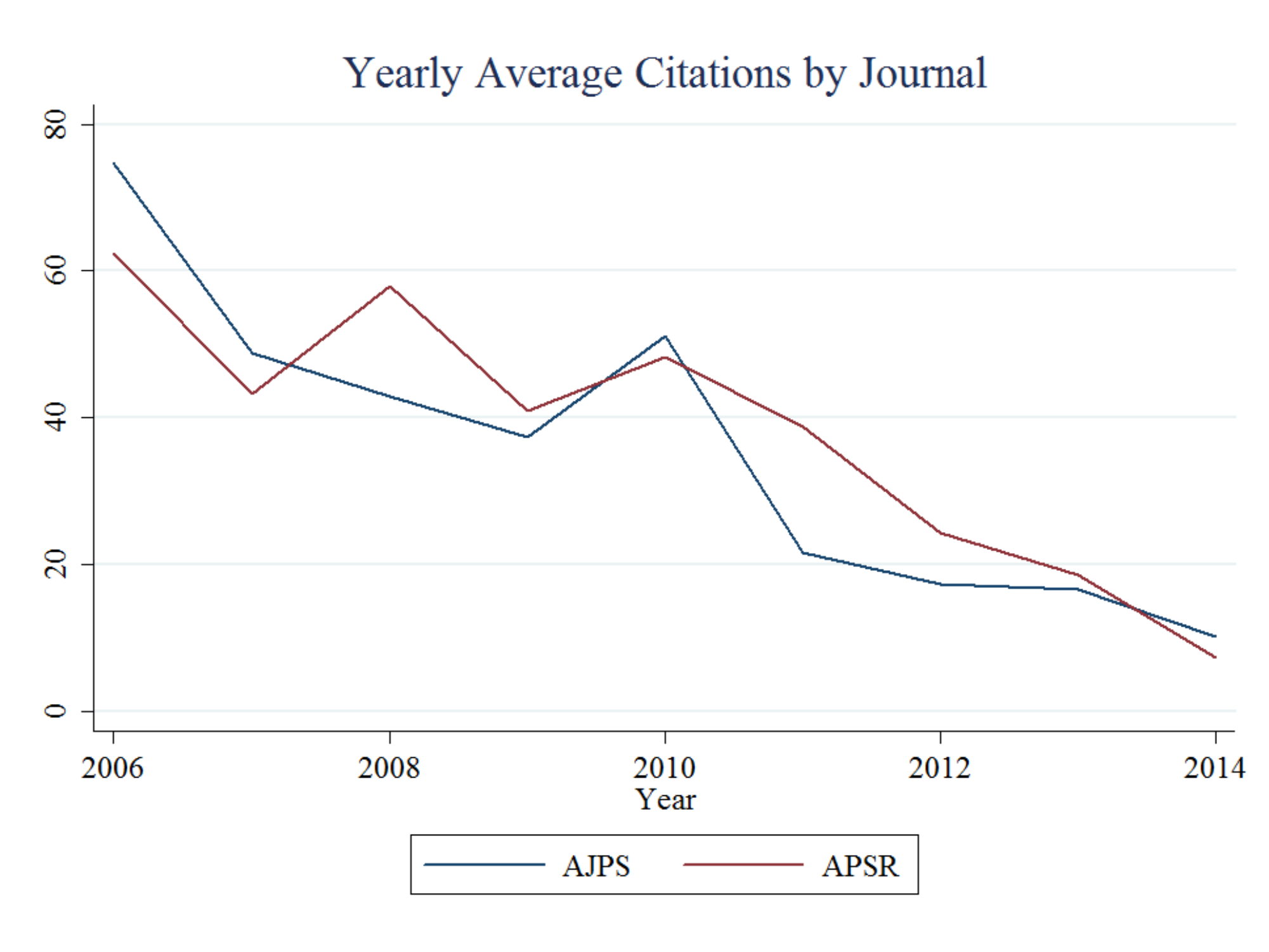
$$availability_i = \alpha_1 + \beta_1 AJPS_i + \beta_2 Post2010_i + \beta_3 Post2012_i \quad (1)$$
$$+ \beta_4 AJPS * Post2010_i + \beta_5 AJPS * Post2012_i$$
$$+ g_1(Time) + h_1(Year) + \nu_i$$

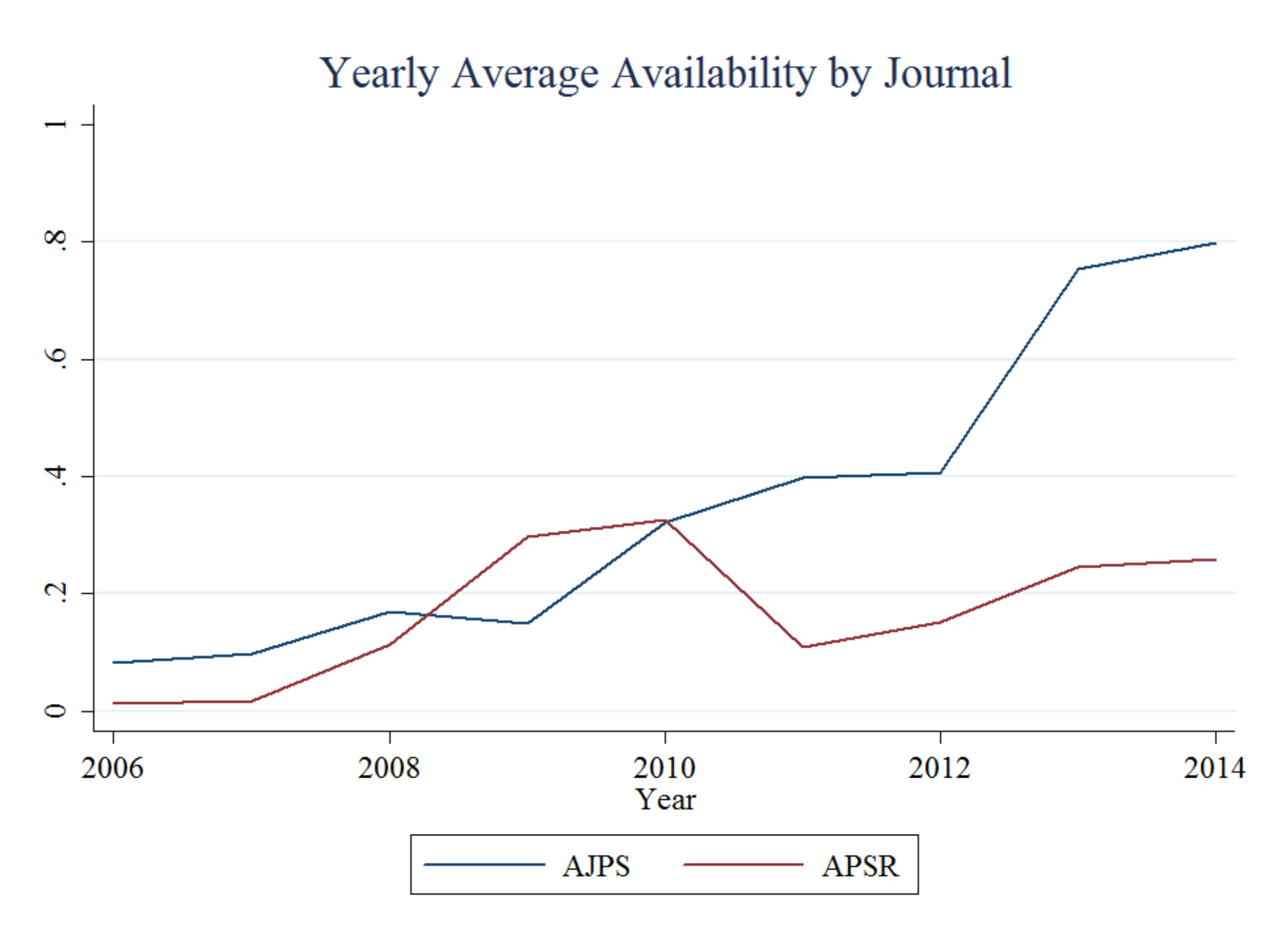
$$citations_i = \alpha_2 + \eta_1 AJPS_i + \eta_2 Post2010_i + \eta_3 Post2012_i \quad (2)$$
$$+ \eta_4 \hat{availability}_i + g_2(Time) + h_2(Year) + u_i$$

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
80
60
40
20
0
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	-3.311	13.89***	8.126**	11.40***
	(3.188)	(3.245)	(3.288)	(3.620)
AJPS		-5.457**	-9.423***	-9.888***
		(2.766)	(2.766)	(3.305)
Months since Pub'd		2.200**	1.892*	2.385*
		(1.112)	(1.090)	(1.322)
Months since Pub'd^2		-0.0198	-0.0170	-0.0215
		(0.0144)	(0.0141)	(0.0172)
Months since Pub'd^3		7.59e-05	6.93e-05	8.55e-05
		(5.80e-05)	(5.68e-05)	(6.93e-05)
No Data in Article			-22.79***	
			(3.428)	
Constant	35.79***	-53.84**	-35.77	-55.13*
	(1.711)	(26.37)	(25.95)	(31.49)
Observations	941	941	938	741
R-squared	0.001	0.164	0.203	0.194
Sample	All	All	All	Data-Only

Standard errors in parentheses

Naive OLS Regression

VARIABLES	(1) Citations	(2) Citations	(3) Citations	(4) Citations
Data and Code Available	-3.311	13.89***	8.126**	11.40***
	(3.188)	(3.245)	(3.288)	(3.620)
AJPS		-5.457**	-9.423***	-9.888***
		(2.766)	(2.766)	(3.305)
Months since Pub'd		2.200**	1.892*	2.385*
		(1.112)	(1.090)	(1.322)
No Data in Article			-22.79***	
			(3.428)	
Observations	941	941	938	741
R-squared	0.001	0.164	0.203	0.194
Sample	All	All	All	Data-Only

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Regressions include constant, squared and cubed months since publication.

Naive OLS Regression

VARIABLES	(1) Ln(Cites+1)	(2) Ln(Cites+1)	(3) Ln(Cites+1)	(4) Ln(Cites+1)
Data and Code Available	-0.0486	0.463***	0.239***	0.278***
	(0.0815)	(0.0795)	(0.0769)	(0.0779)
AJPS		-0.139**	-0.299***	-0.270***
		(0.0678)	(0.0647)	(0.0712)
Months since Pub'd		0.0899***	0.0805***	0.0909***
		(0.0273)	(0.0255)	(0.0285)
No Data in Article			-0.865***	
			(0.0801)	
Observations	941	941	938	741
R-squared	0.000	0.232	0.321	0.281
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	-0.0121	-0.0300
	(0.0343)	(0.0323)	(0.0464)
Post-Oct 2010	-0.225***	-0.147	-0.245**
	(0.0788)	(0.0905)	(0.0981)
Post-July 2012	-0.0955	-0.303***	-0.107
	(0.0787)	(0.0985)	(0.0970)
Post-2010 with Data		-0.0478	
		(0.0899)	
Post-2012 with Data		0.234**	
		(0.102)	
Months since Pub'd	-0.00568	-0.0132	-0.0201
	(0.0131)	(0.0125)	(0.0162)
No Data in Article		-0.165***	
		(0.0389)	
AJPS post-2010 Policy	0.205***		0.219***
	(0.0664)		(0.0839)
AJPS post-2012 Policy	0.268***		0.289***
	(0.0738)		(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

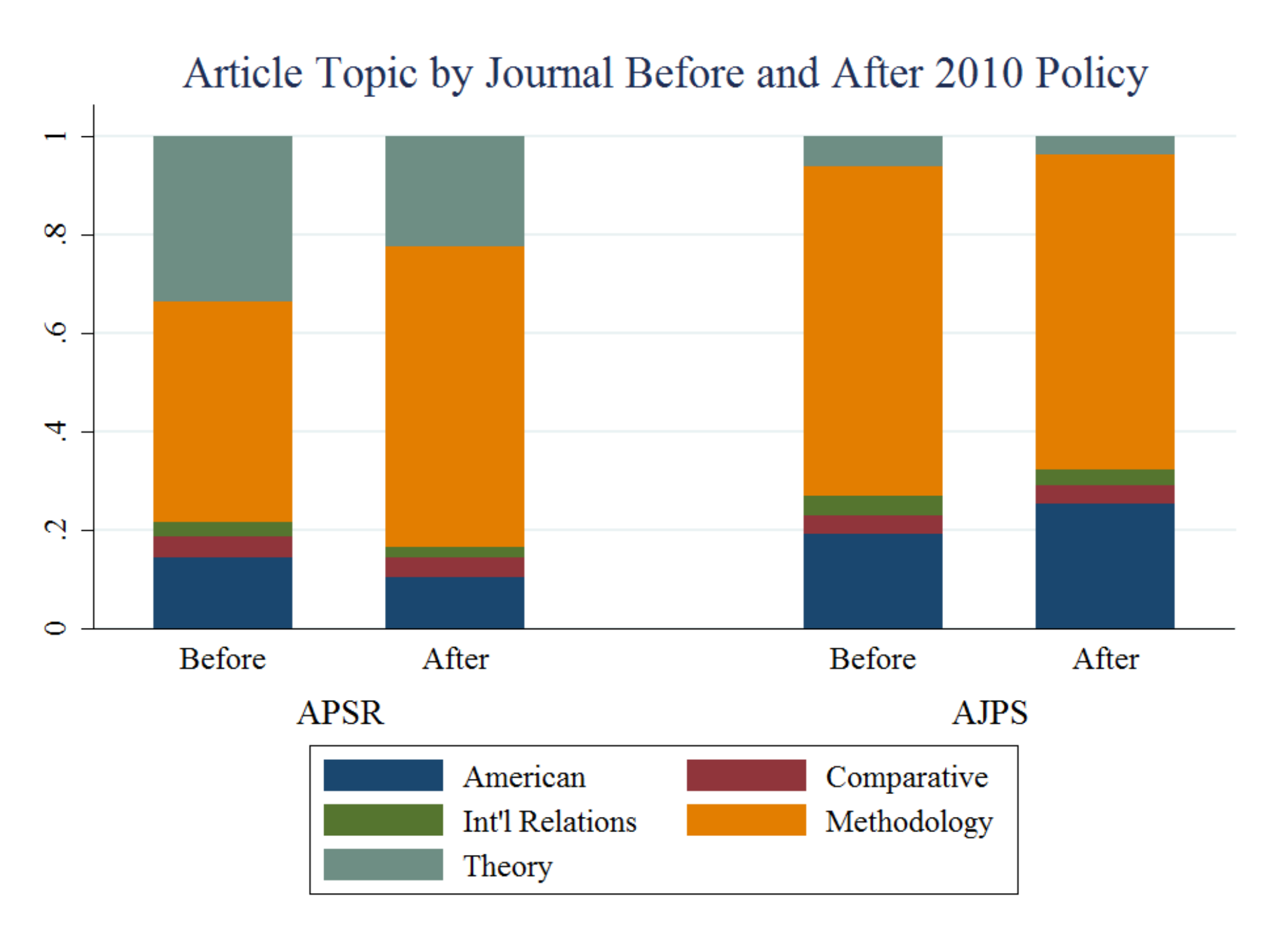
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

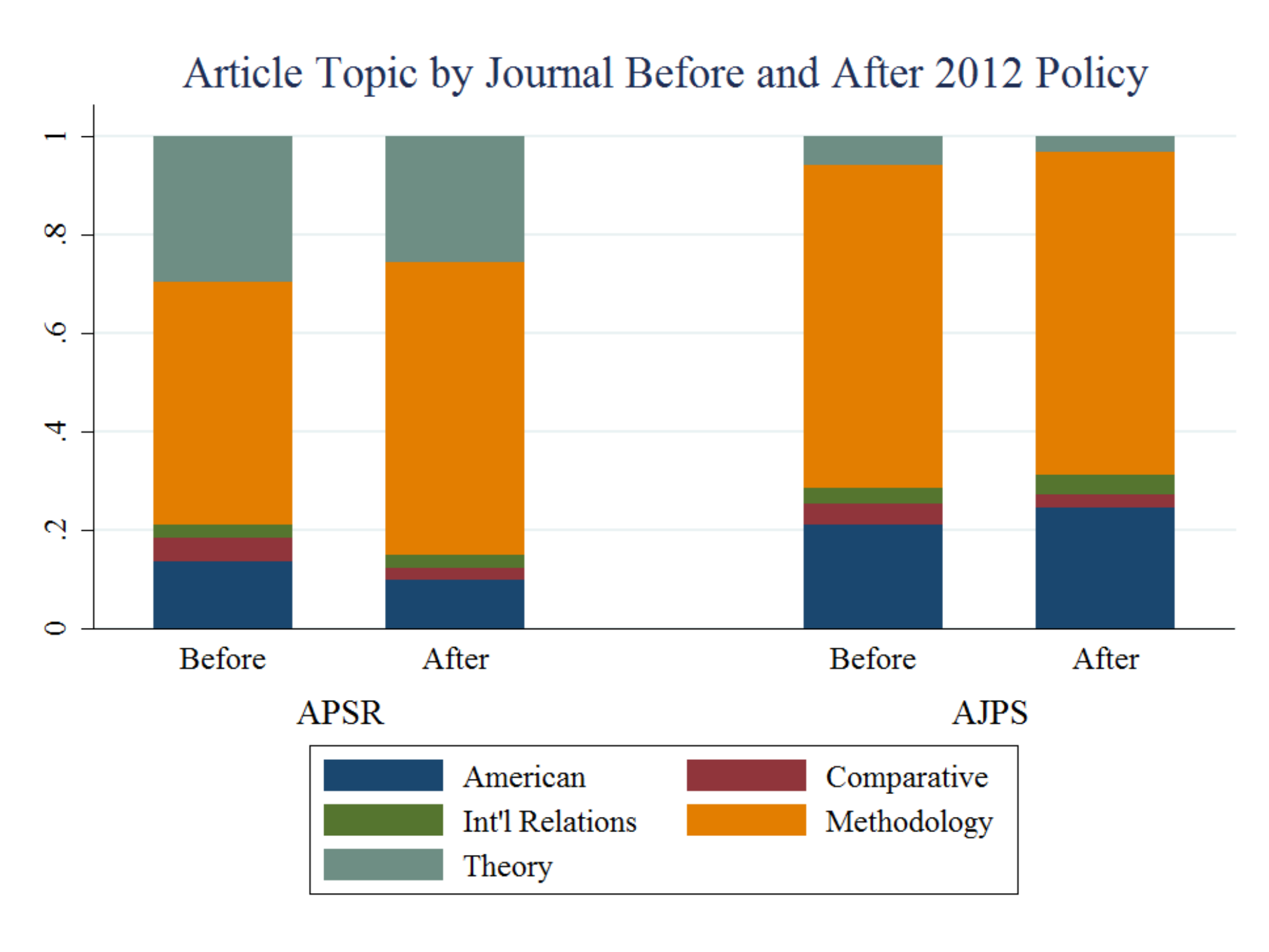
2SLS Regression

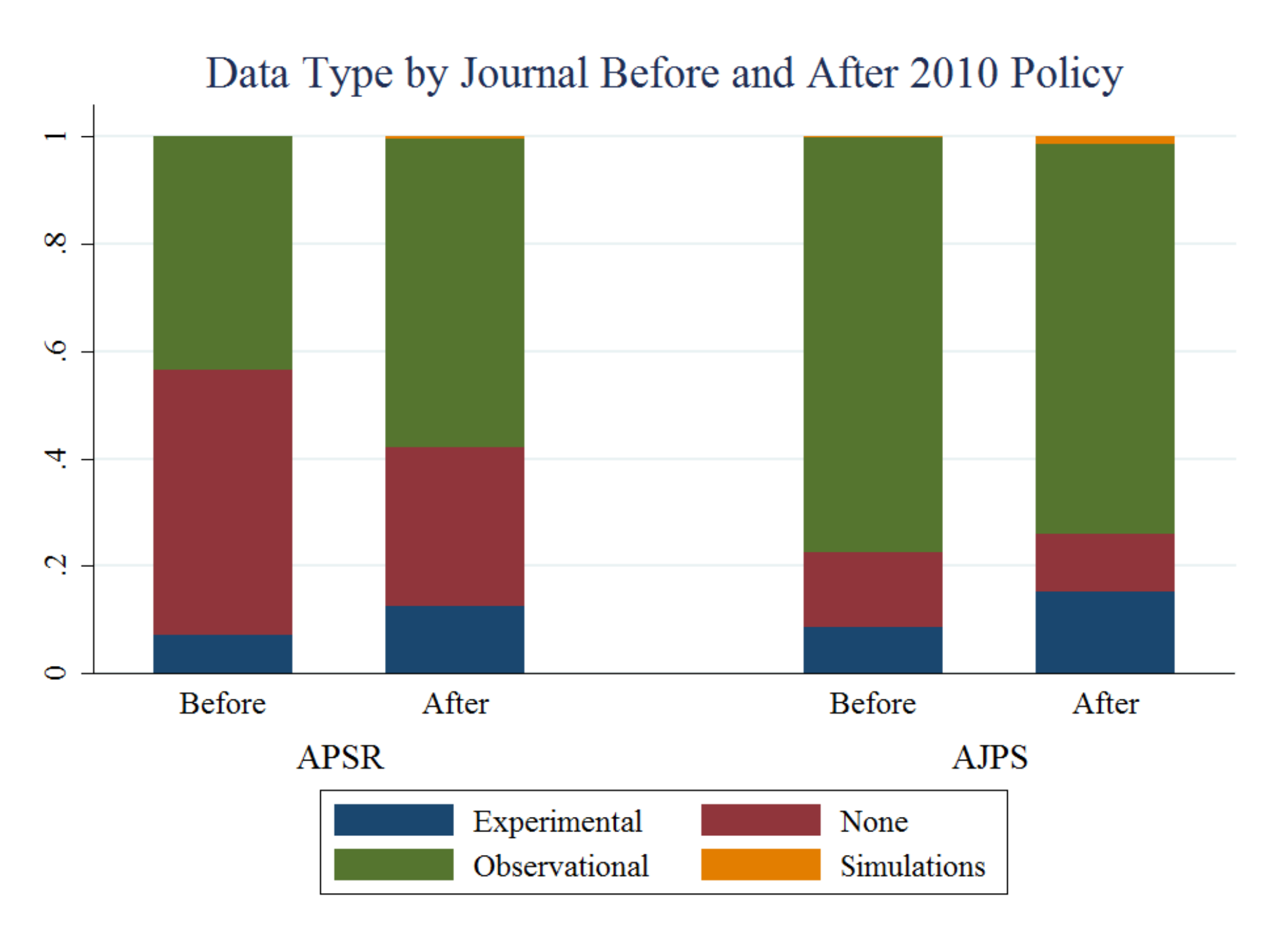
VARIABLES	(1) 2SLS	(2) 2SLS	(3) 2SLS
Data and Code Available	-4.783	7.214	7.322
	(13.02)	(13.18)	(14.52)
AJPS	-2.126	-9.613***	-9.564**
	(3.791)	(3.294)	(4.116)
Post-Oct 2010	-22.30***	-3.056	-21.77**
	(7.566)	(9.932)	(8.670)
Post-July 2012	-7.962	1.967	-9.500
	(6.915)	(11.22)	(7.725)
Post-2010 with Data		-16.61**	
		(8.474)	
Post-2012 with Data		-11.37	
		(10.86)	
Months since Pub'd	3.863***	3.443**	4.196**
	(1.391)	(1.340)	(1.651)
No Data in Article		-33.18***	
		(4.680)	
Observations	941	938	741
R-squared	0.143	0.221	0.200
Sample	All	IV=Data-Only	Data-Only

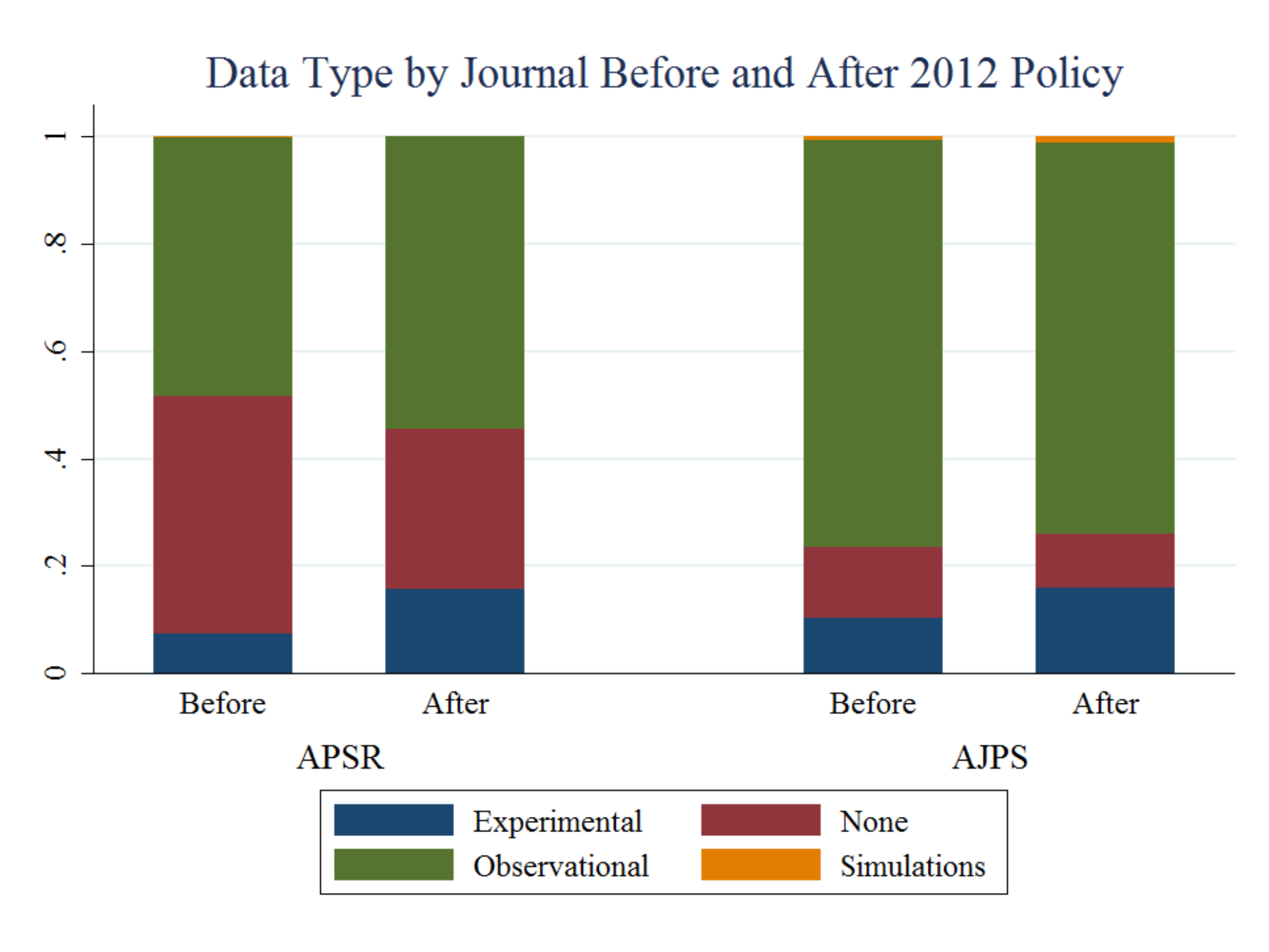
2SLS Regression of ln(citations+1)

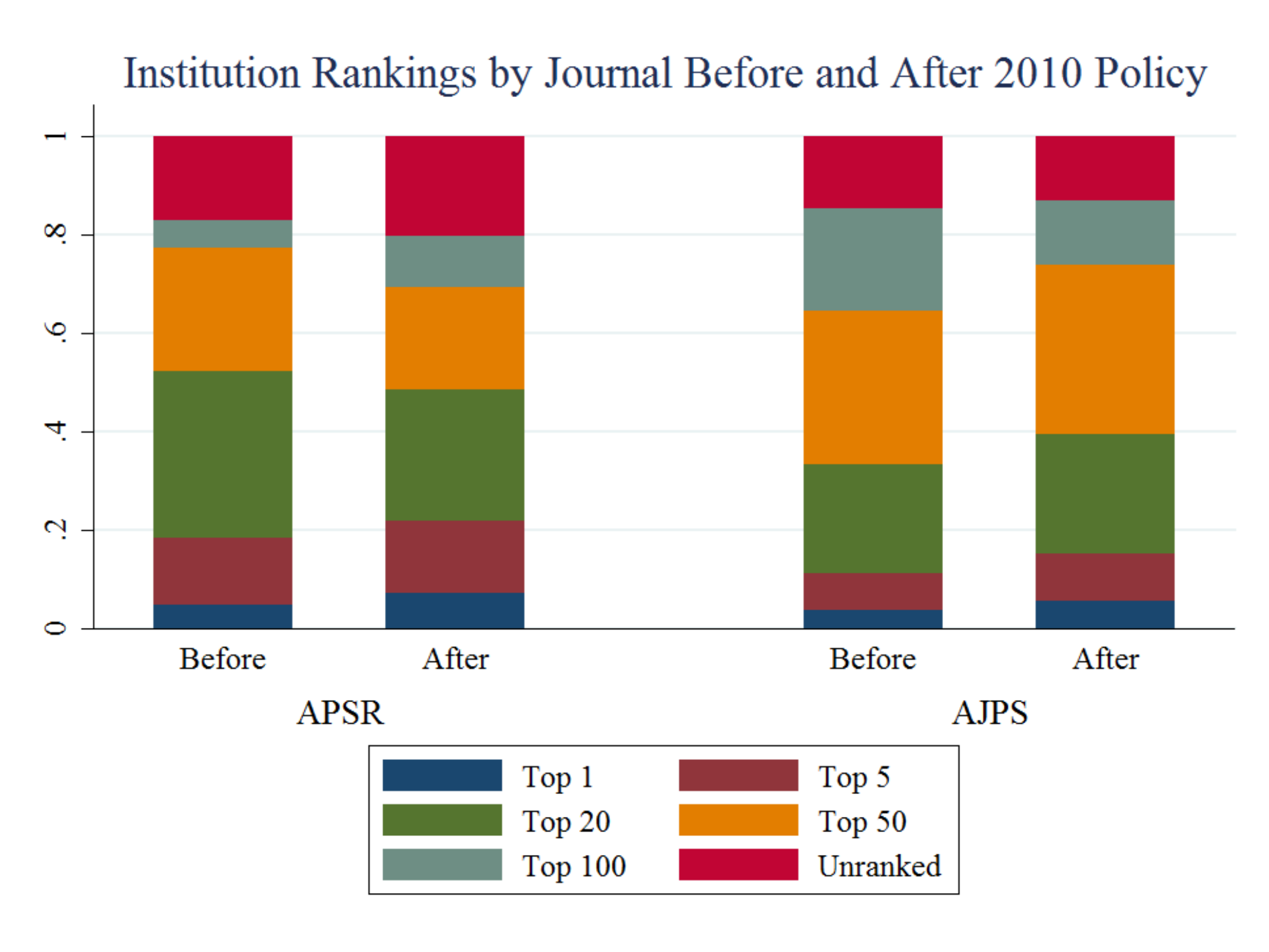
VARIABLES	(1) 2SLS-Log	(2) 2SLS-Log	(3) 2SLS-Log
Data and Code Available	0.0419	0.228	0.198
	(0.320)	(0.311)	(0.314)
AJPS	-0.0564	-0.299***	-0.259***
	(0.0931)	(0.0777)	(0.0889)
Post-Oct 2010	-0.150	0.224	-0.175
	(0.186)	(0.234)	(0.187)
Post-July 2012	-0.0833	-0.0123	-0.110
	(0.170)	(0.265)	(0.167)
Post-2010 with Data		-0.244	
		(0.200)	
Post-2012 with Data		-0.0523	
		(0.256)	
Months since Pub'd	0.0937***	0.0758**	0.102***
	(0.0341)	(0.0316)	(0.0357)
No Data in Article		-0.995***	
		(0.110)	
Observations	941	938	741
R-squared	0.209	0.324	0.281
Sample	All	IV=Data-Only	Data-Only

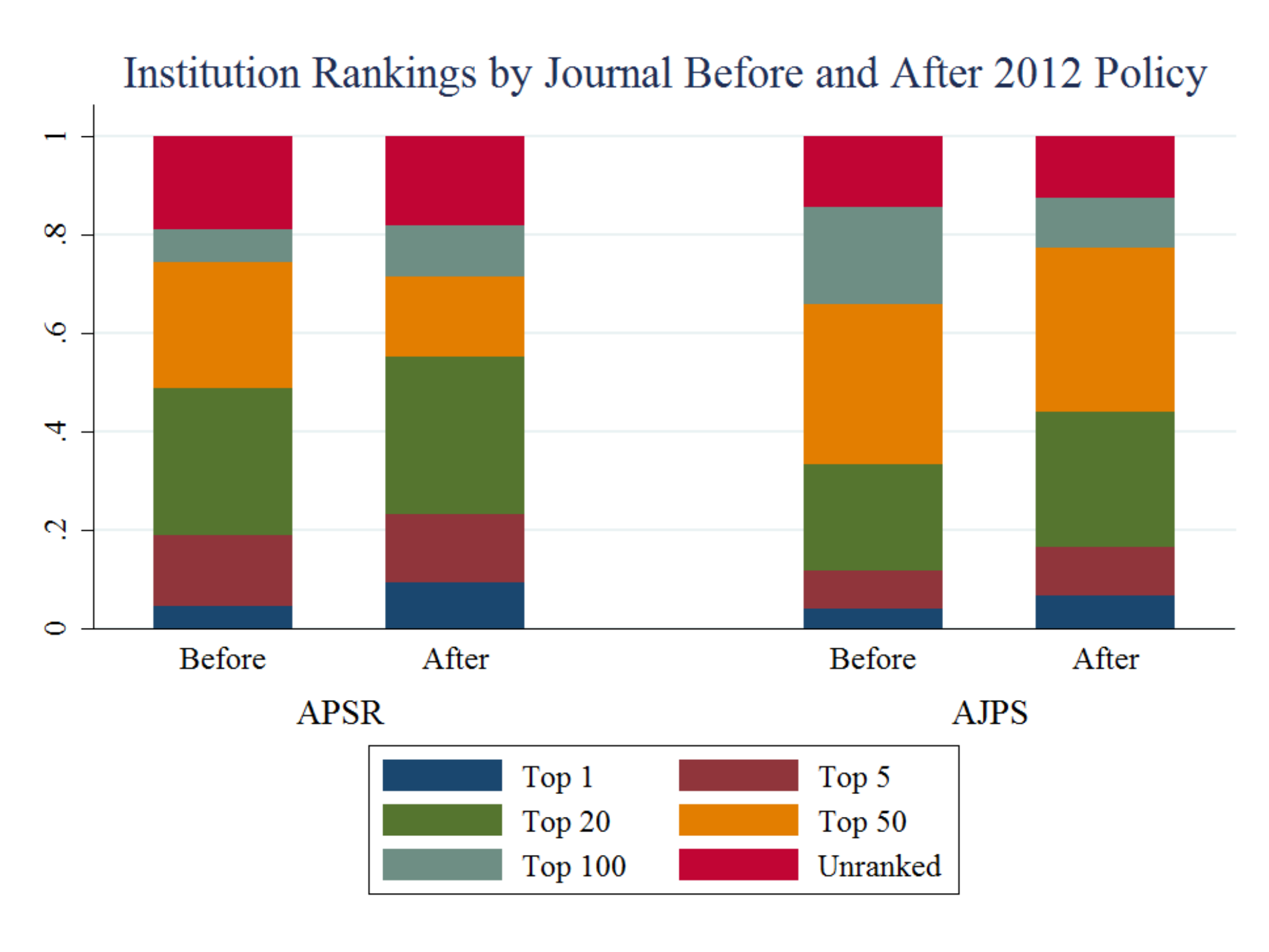

Article Topic by Journal Before and After 2010 Policy
0
.2
.4
.6
.8
1
Before
After
Before
After
APSR
AJPS
American
Comparative
Int'l Relations
Methodology
Theory


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	-0.0993	9.98e-05	0.160*
	(0.0827)	(0.0928)	(0.0709)	(0.0725)	(0.0595)	(0.0887)
AJPS post-2012 Policy	-0.00687	0.00991	-0.105	0.0919	0.00406	0.0139
	(0.0891)	(0.100)	(0.0764)	(0.0782)	(0.0645)	(0.0961)
AJPS	0.0287	0.0254	-0.0419	0.0376	-0.0594*	-0.201***
	(0.0457)	(0.0513)	(0.0392)	(0.0401)	(0.0330)	(0.0492)
Post-Oct 2010	-0.0168	0.00684	-0.000582	-0.000418	0.0267	-0.0949
	(0.0966)	(0.109)	(0.0828)	(0.0847)	(0.0694)	(0.103)
Post-July 2012	-0.0513	0.0786	0.163**	-0.153*	0.0285	0.0744
	(0.0956)	(0.107)	(0.0820)	(0.0838)	(0.0690)	(0.103)
Months since Pub'd	-0.0251	0.0331*	-0.00167	-0.00439	0.00289	-0.00648
	(0.0160)	(0.0179)	(0.0137)	(0.0140)	(0.0115)	(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