False Positives in Policy Research

Sean Tanner
University of California, Berkeley

APPAM Spring Conference
4-11-15

Prepared for APPAM's 2015 spring conference. Please do not cite without permission.

Credibility Crisis in Social Science

- Recent wave of interest in long-standing concerns over false-positives

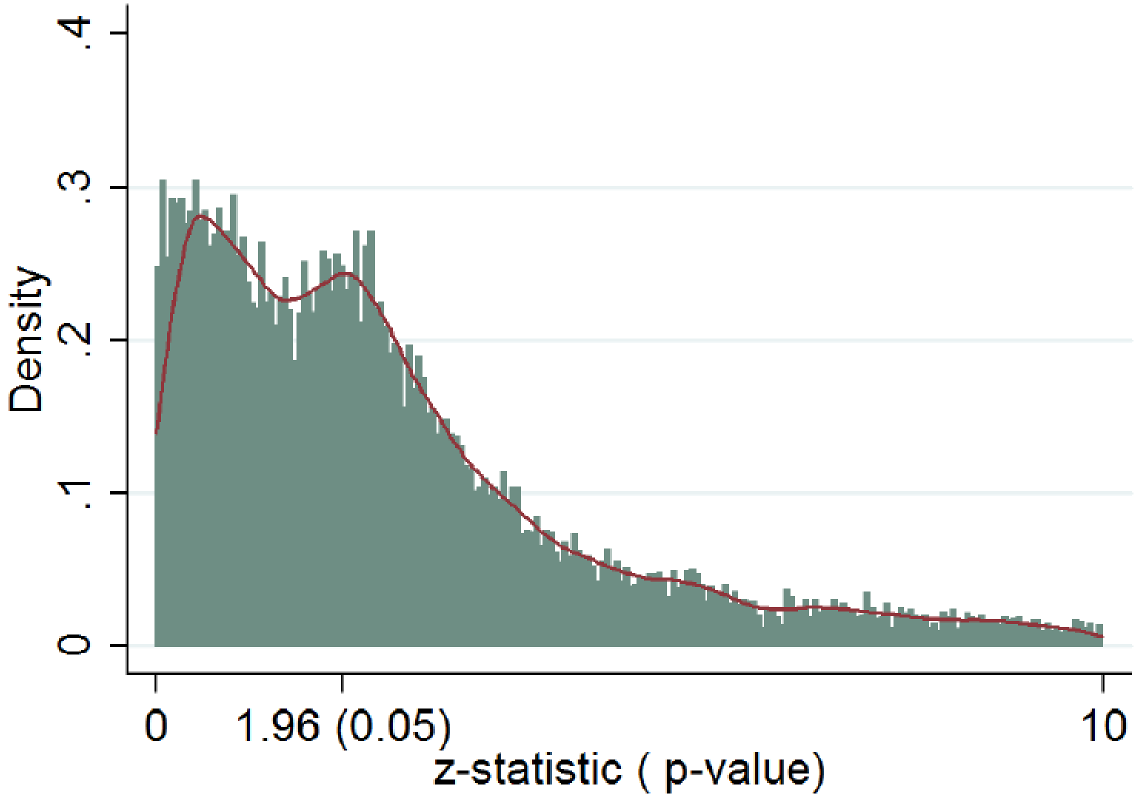
- False positive = reported effect when the truth is no effect

- Despite rigorous methods (RCT, RD, IV), many findings are fragile at best
  - Outright fraud/fabricated data
  - Questionable sample restrictions/specifications

Three Reasons for False Positives

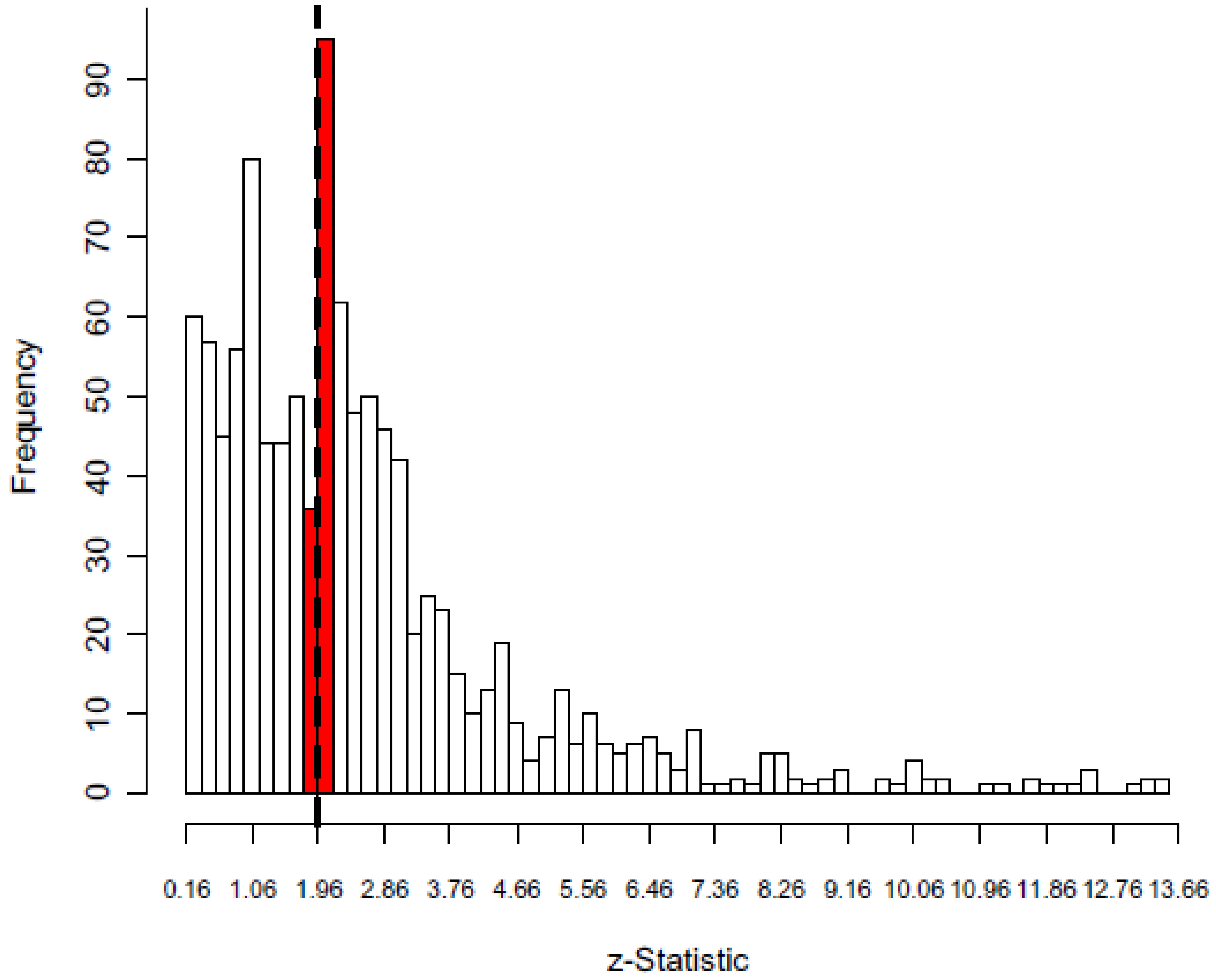
- Sampling Variance
  - Valid inferential technique, but “bad draw”
- File-drawer
  - Whole studies left unpublished due to null findings
- P-hacking (“Specification Search” or “Massaging the Data”)
  - Altering specification until a significant effect is found
  - If p-hacking exists, p-values and test stats cluster (.1, .05)

Clustering in Economics



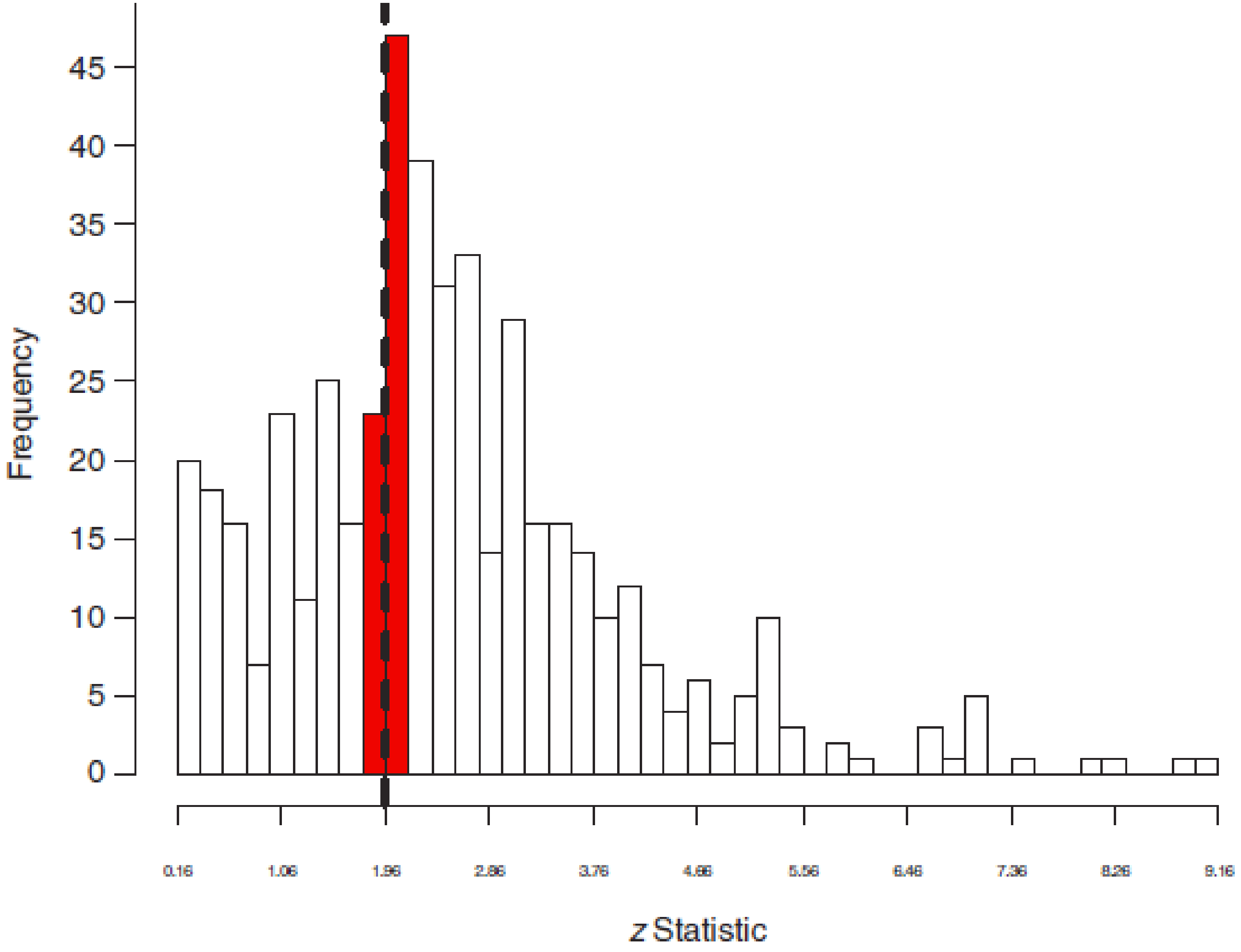
Brodeur, Abel; Lé, Mathias; Sangnier, Marc; Zylberberg, Yanos (2013) : Star Wars: The empirics strike back, Discussion Paper Series, Forschungsinstitut zur Zukunft der Arbeit, No. 7268
http://hdl.handle.net/10419/71700
50,000 tests published between 2005 and 2011 in the *AER*, *JPE*, and *QJE*

Clustering in Political Science



Source: Gerber and Malhotra, 2008a. Data from *APSR* & *AJPS*

Clustering in Sociology



Source: Gerber and Malhotra, 2008b. Data from *American Journal of Sociology* & *The Sociological Quarterly*

Contributions of this research

- Formally models p-hacking
  - Only “significant” p-values
  - Statistically independent tests (one per article)
- Focuses on rigorous, policy-relevant work

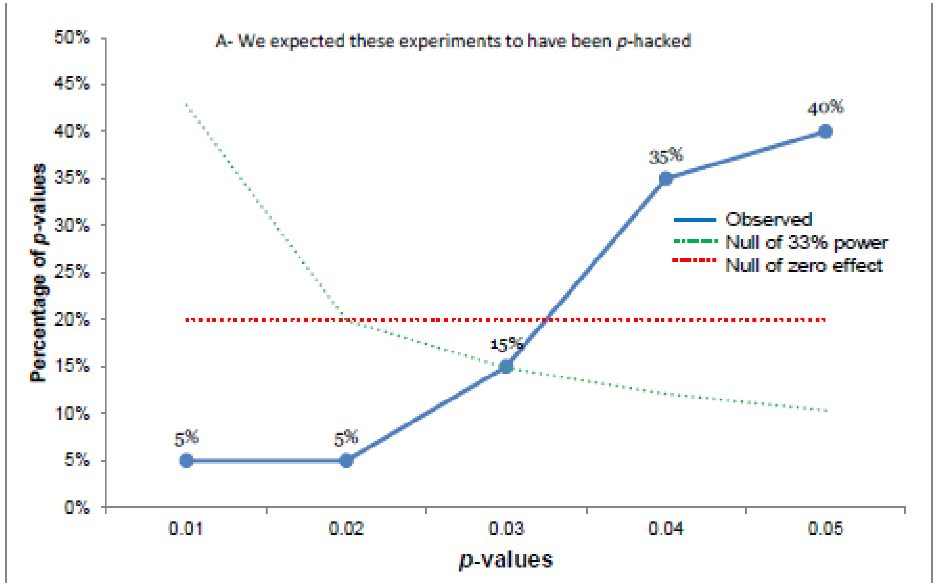
How to detect p-hacking

- P-curve (Simonsohn, Nelson, Simmons, 2014)

- Distribution of observed p-values

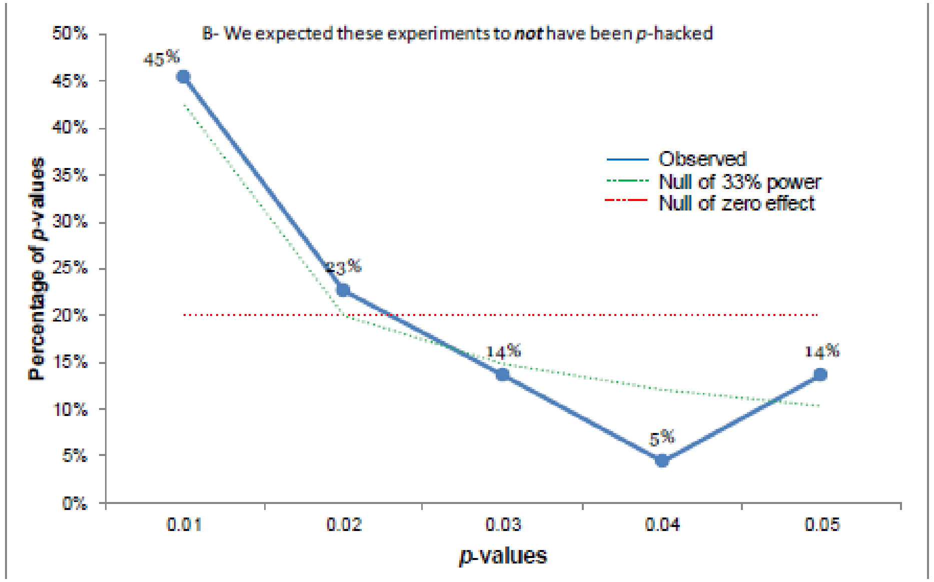
- Should only be uniform (flat) or right-skewed

P-Curve



Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-Curve: A Key to the File-Drawer. *Journal of experimental psychology.*

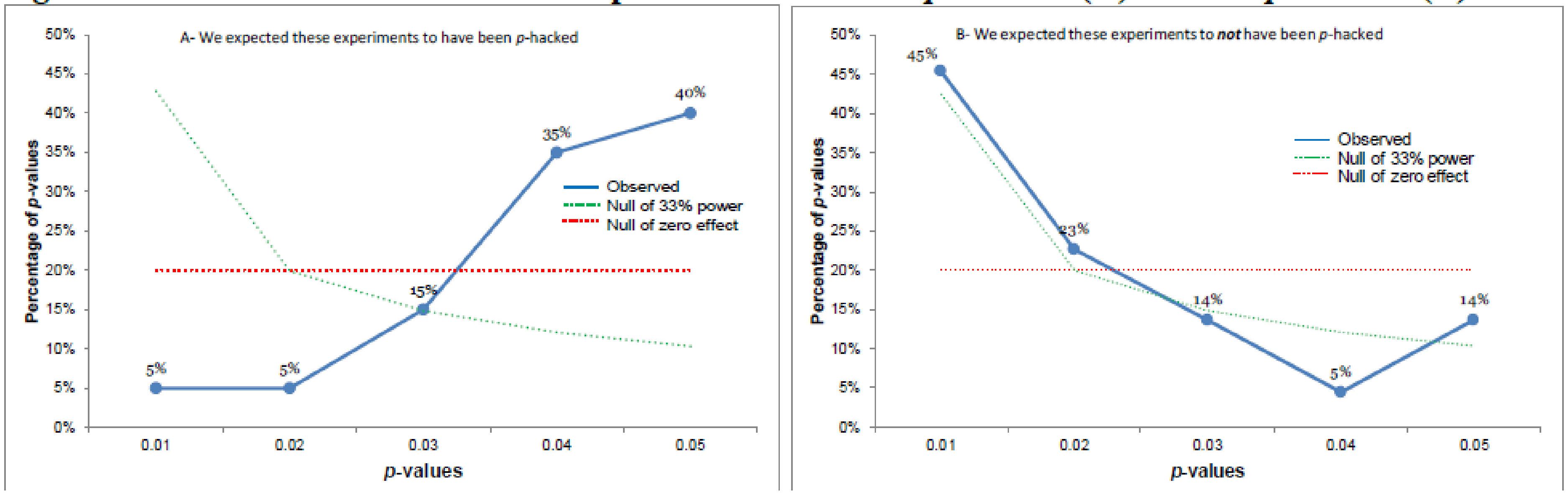
P-Curve



Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). P-Curve: A Key to the File-Drawer. *Journal of experimental psychology.*

P-Curve

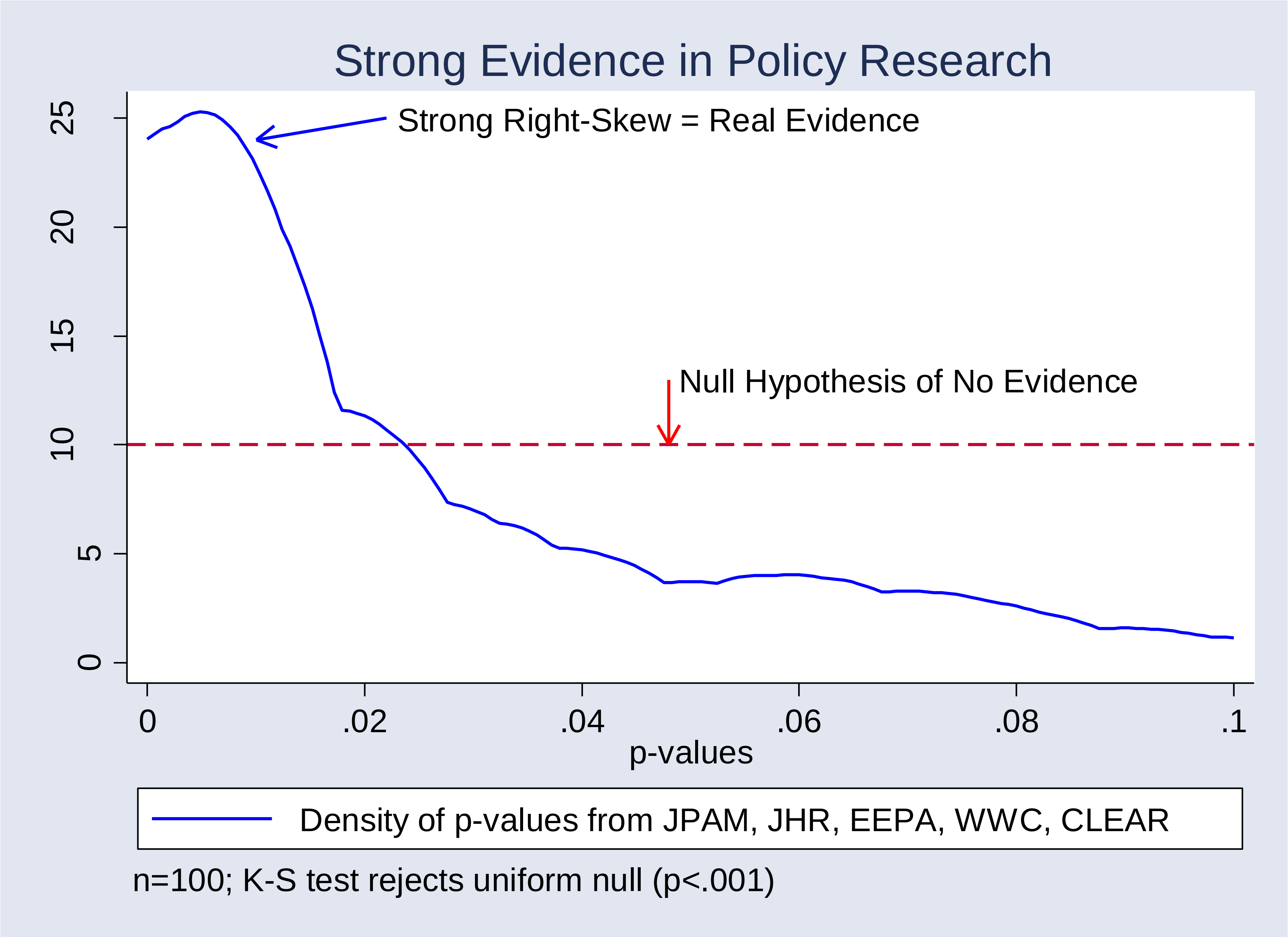
Figure 3. ***P*-curves for JPSP studies suspected to have been *p*-hacked (A) and not *p*-hacked (B).**



Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). P-Curve: A Key to the File-Drawer. *Journal of experimental psychology.*

P-hacking in Policy Research?

- What Works Clearinghouse (DoED)
  Clearinghouse of Labor Evaluation and Research (DOL)
- *Journal of Policy Analysis and Management*
  - Two “similar” journals (Reuter, P. & Smith-Ready, J., *JPAM*, 2002)
  - *Journal of Human Resources*
  - *Education Evaluation and Policy Analysis*



Strong Evidence in Policy Research

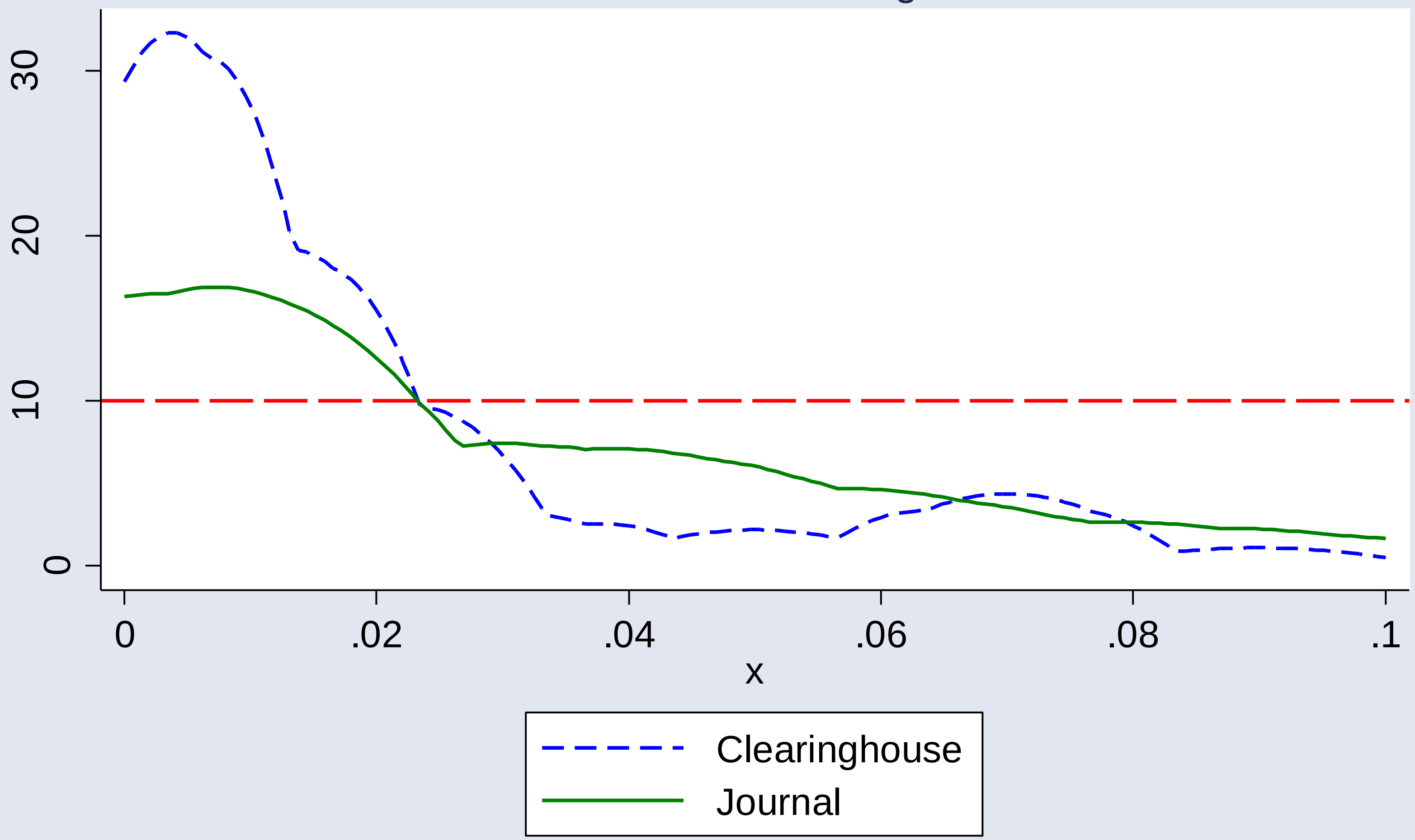
Strong Right-Skew = Real Evidence

Null Hypothesis of No Evidence

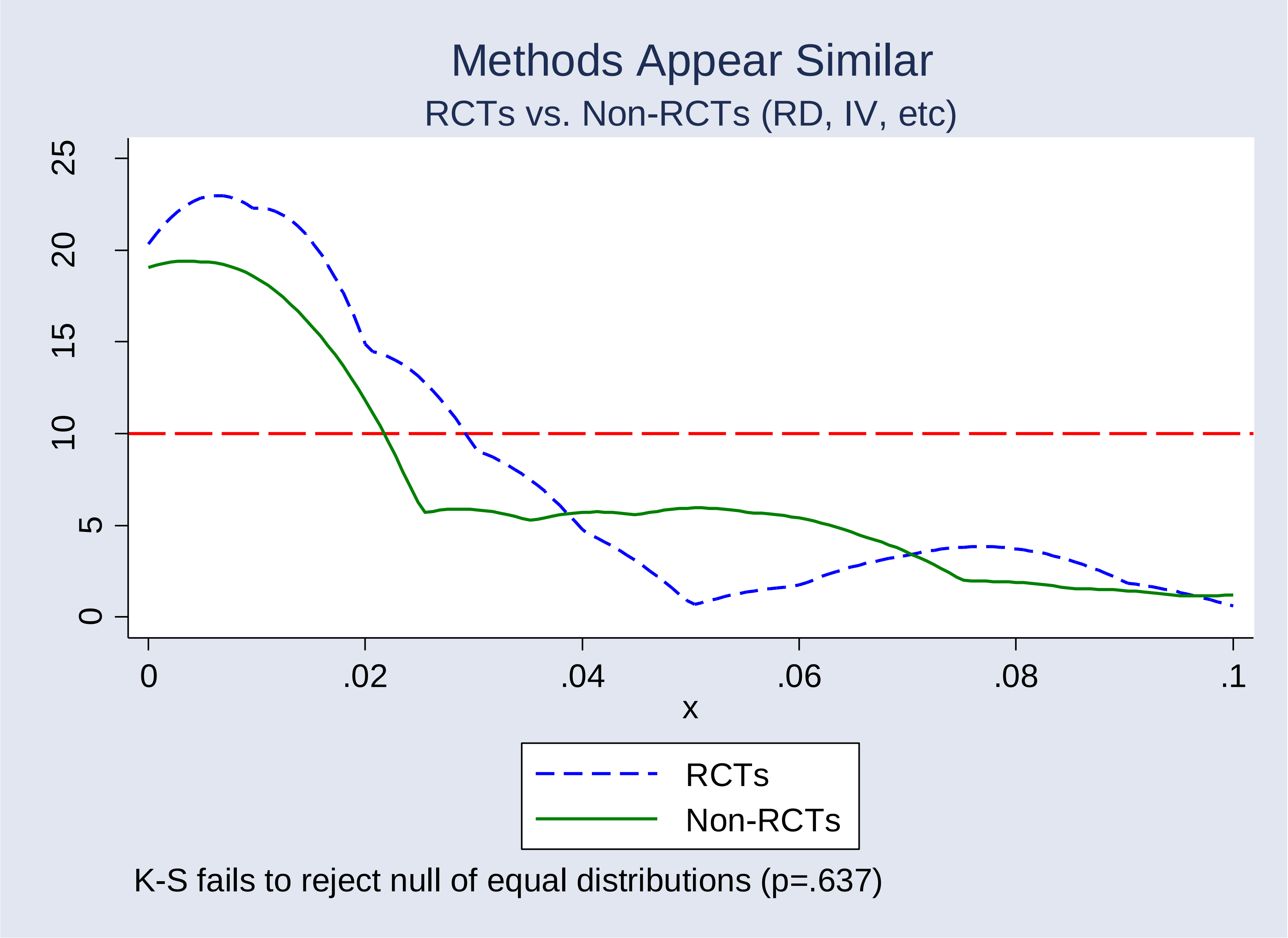
n=100; K-S test rejects uniform null (p<.001)

Stronger Evidence in Clearinghouses

Journal Distribution Less Right-Skewed



K-S fails to reject null of equal distributions (p=.247)



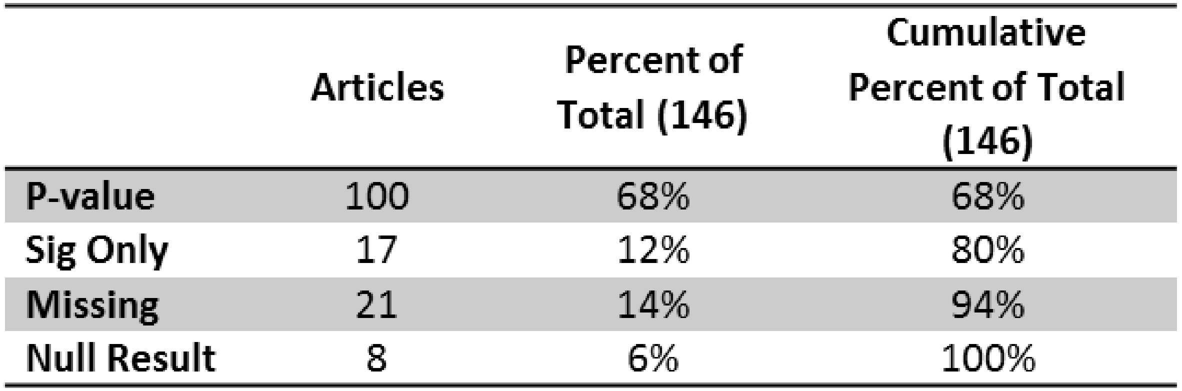
Methods Appear Similar

RCTs vs. Non-RCTs (RD, IV, etc)

K-S fails to reject null of equal distributions (p=.637)

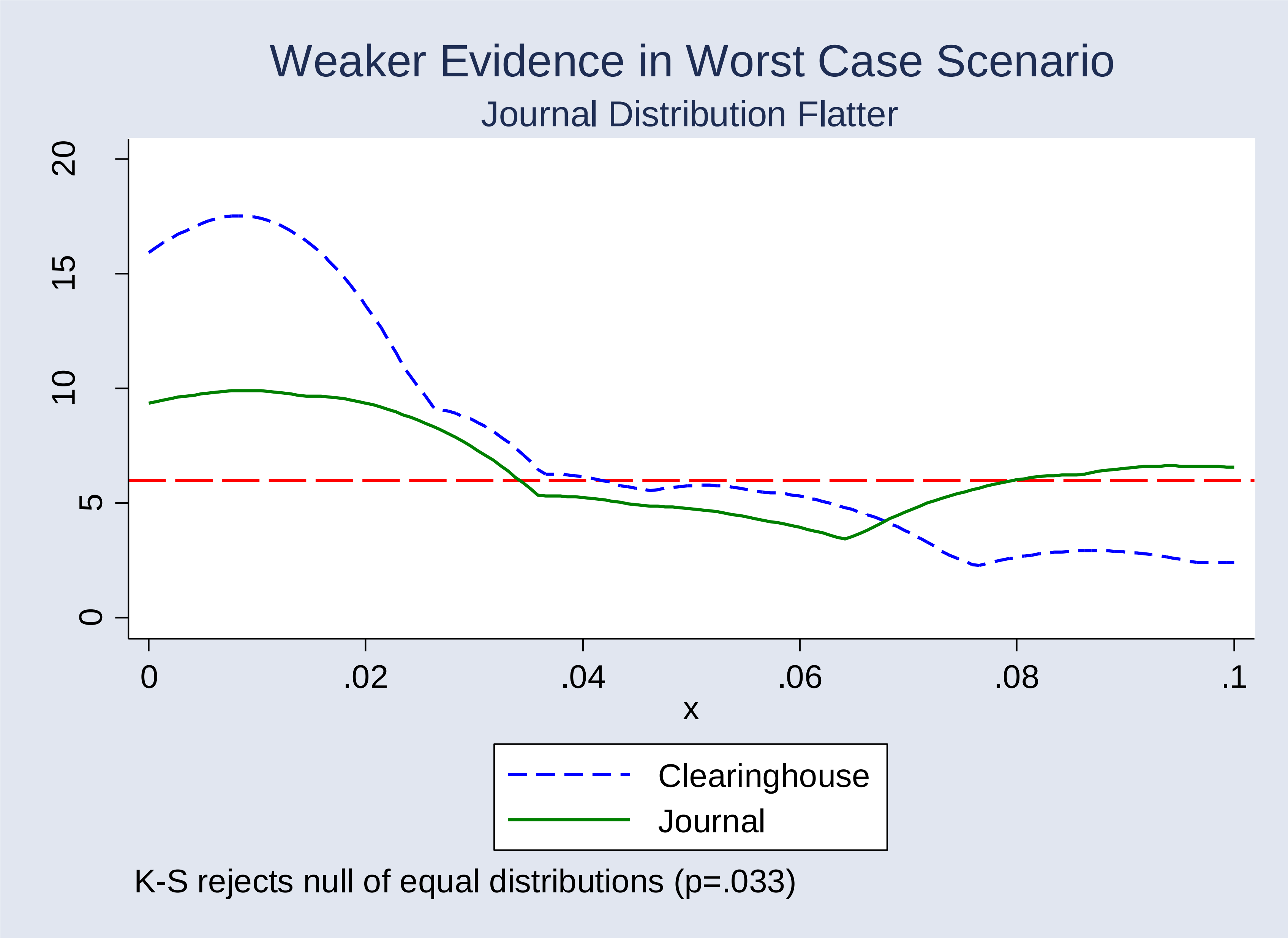
Missing P-values

- Only 68% of p-values available



	Articles	Percent of Total (146)	Cumulative Percent of Total (146)
P-value	100	68%	68%
Sig Only	17	12%	80%
Missing	21	14%	94%
Null Result	8	6%	100%

- Worst case scenario: p-value= sig level or .1 if missing completely



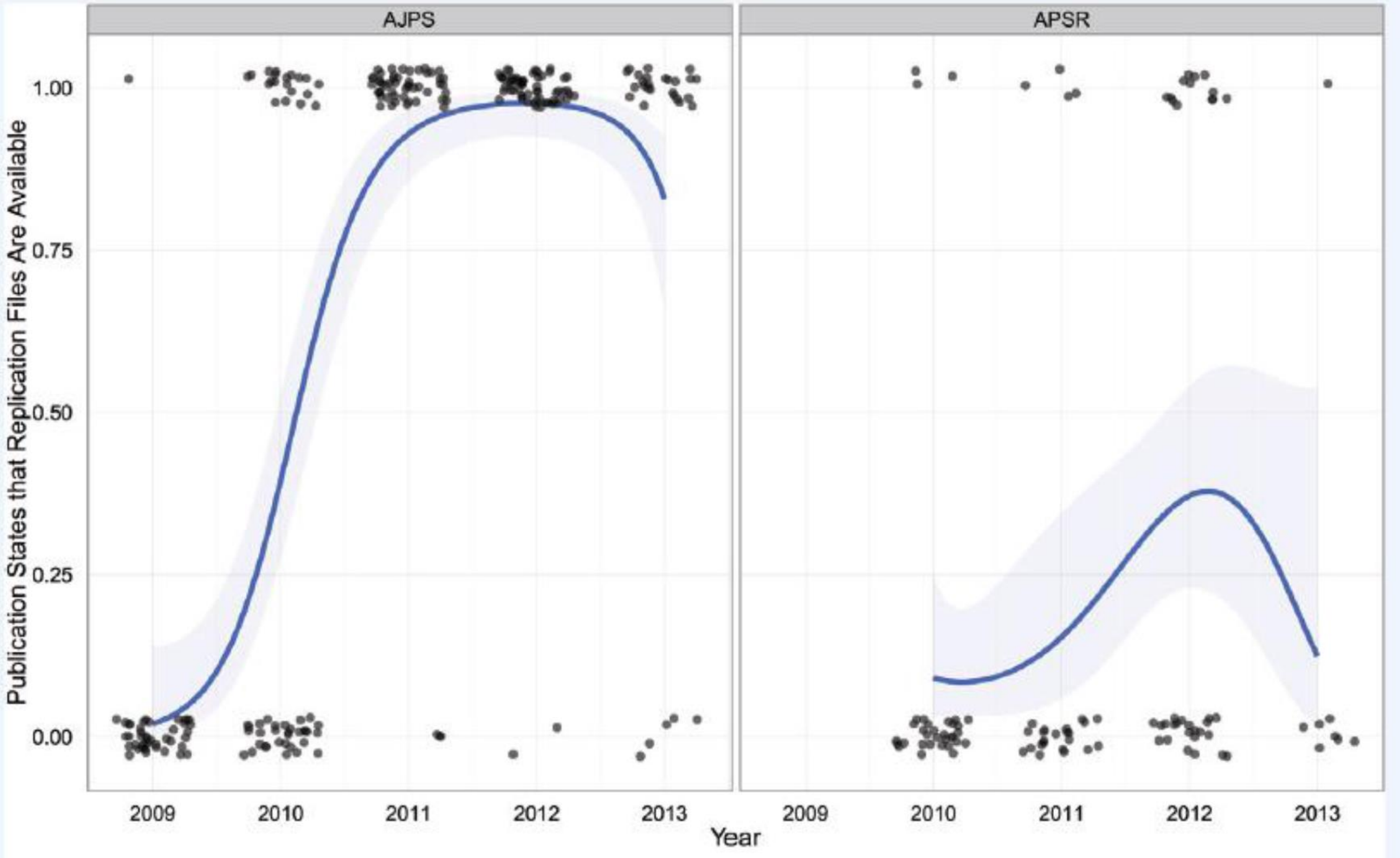
Weaker Evidence in Worst Case Scenario

Journal Distribution Flatter

K-S rejects null of equal distributions (p=.033)

Increased Credibility through Transparency

- Strong evidence in policy research
  - Still unable to evaluate any single study
  - What happens when federal funding is linked to RCT results?
- Three mechanisms for increasing transparency
  1. Registration and pre-analysis plans (PAPs)
  2. Open materials (data & code)
  3. Disclosure


AJPS
APSR
Publication States that Replication Files Are Available
1.00
0.75
0.50
0.25
0.00
2009
2010
2011
2012
2013
2009
2010
2011
2012
2013
Year

Dafoe, 2014

Potential Actions from JPAM & APPAM

- JPAM
  - Endorse principles as other journals have
  - Encourage registration and PAPs
  - Make code and data sharing the default (as AER & AJPS do)
  - Symposium
- APPAM
  - Workshop at fall meeting
  - Reproduction contest for graduate students

Registration and Pre-analysis Plans (PAPs)

- PAPs demarcates ex-ante vs. ex-post analyses
- AEA, 3ie, EGAP now have registries
  - AEA has 297 trials; 61 have PAPs
  - 3ie has 40 trials
  - EGAP has 121 trials, 41 have PAPs
- Required by law in clinical trials
- The PAP for this research: Tanner (2015)

Open Data and Materials

- Helps replication, minimizes threat of fraud, advances science
- Endorsed by *Nature*, *Science*, *AER*, NSF, NIH, Royal Society
- Center for Open Science & Dataverse assist researchers

Disclosure

- Partially integrated through online appendices
- Standard disclosure checklist?
  - CONSORT for clinical trials
- Finkelstein et al (2012) used ^ to denote supplemental hypotheses

- Berkeley Initiative for Transparency in Social Science (BITSS) http://bitss.org/

References

- Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2013). *Star Wars: The Empirics Strike Back* (No. Discussion Paper Series, Forschungsinstitut zur Zukunft der Arbeit, No. 7268).
- Dafoe, A. (2014). Science Deserves Better: The Imperative to Share Complete Replication Files. *PS, Political Science & Politics*, *47*(1), 60–66.
- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., … Group, O. H. S. (2012). The Oregon Health Insurance Experiment: Evidence From The First Year. *Quarterly Journal of Economics*, *127*(August (3)), 1057–1106.
- Gerber, A., & Malhotra, N. (2008a). Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science*, *3*(3), 313–326.
- Gerber, A., & Malhotra, N. (2008b). Publication Bias in Empirical Sociological Research. *Sociological Methods & Research*, *37*(1), 3 –30.
- Reuter, P., & Smith-Ready, J. (2002). Assessing JPAM after 20 Years. *Journal of Policy Analysis and Management*, *21*(3), 339–353.
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology. General*, *143*(2), 534–47.
- Tanner, S. (2015). False Positives and Selective Reporting in Policy. *Observational Studies*, *1*(1), 18–29.