Transparency, Reproducibility, and the Credibility of Economics Research

Garret Christensen and Edward Miguel

Journal of Economic Literature 2018, 56(3), 920–980

James J. Heckman

THE UNIVERSITY OF CHICAGO

Econ 312, Spring 2023

1. Introduction

- Openness and transparency have long been considered key pillars of the scientific ethos (Merton 1973).

- Yet there is growing awareness that current research practices often deviate from this ideal, and can sometimes produce misleading bodies of evidence (Miguel et al. 2014).

2. Evidence on Problems with the Current Body of Research

2.1 A Model for Understanding the Issues

- Specifically, the model estimates the positive predictive value (PPV) of research, or the likelihood that a claimed empirical relationship is actually true, under various assumptions.

- Define R_i as the ratio of true relationships to no relationships commonly tested in a research field i (e.g., development economics).

- Prior to a study being undertaken, the probability that a true relationship exists is thus R_i /(R_i +1).

- Using the usual notation for statistical power of the test (1 – β) and statistical significance level (α), the PPV in research field *i* is given by

$$(1) \qquad PPV_i = \frac{(1-\beta)R_i}{(1-\beta)R_i + \alpha}.$$

- Clearly, the better powered the study, and the stricter the statistical significance level, the closer the PPV is to one, in which case false positives are largely eliminated.

- At the usual significance level of α = 0.05 and in the case of a well-powered study (1 − β = 0.80) in a literature in which half of all hypotheses are thought to be true ex ante (R_i /(R_i +1)= 0.5), the PPV is relatively high at 94 percent, a level that would not seem likely to threaten the validity of research in a particular economics subfield.

- What are true α and β?

- This concern, and those discussed next, are all exacerbated by bias in the publication process.

- Denoted by u, researcher bias is defined as the probability that a researcher presents a non- finding as a true finding, for reasons other than chance variation in the data.

- This researcher bias could take many forms, including any combination of specification searching, data manipulation, selective reporting, and even outright fraud.

Extending the above framework to incorporate the researcher bias term (u_i) in field i leads to the following expression:

(2)

$$PPV_i = \frac{(1-\beta)R_i + u_i\beta R_i}{(1-\beta)R_i + \alpha + u_i\beta R_i + u_i(1-\alpha)}.$$

- In table 1 (a reproduction of table 4 from Ioannidis 2005), we present a range of parameter values and the resulting PPV.

- First off, literatures characterized by statistically underpowered (i.e., small $1 - \beta$) studies are likely to have many false positives.

- Second, the hotter a research field, with more teams (n_i) actively running tests and higher stakes around the findings, the more likely it is that findings are false positives.

- Third, the greater the flexibility in research design, definitions, outcome measures, and analytical approaches in a field, the less likely the research findings are to be true, again due to a combination of multiple testing concerns and author bias.

TABLE 1
POSITIVE PREDICTIVE VALUE (PPV) OF RESEARCH FINDINGS FOR VARIOUS COMBINATIONS OF POWER ($1 - \beta$), RATIO OF TRUE TO NOT-TRUE RELATIONSHIPS (R), AND RESEARCHER BIAS (u)

$1 - \beta$	R	u	Practical example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

Notes: The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study. RCT, randomized controlled trial.
Source: Reproduced from table 4 of Ioannidis (2005). DOI: 10.1371/journal.pmed.0020124.t004

2.2 Publication Bias

- The term “file drawer problem” was coined decades ago (Rosenthal 1979) to describe this problem of results that are missing from a body of research evidence.

- Important recent research by Franco, Malhotra, and Simonovits (2014) affirms the importance of this issue in practice in contemporary social science research.

- They document that a large share of empirical analyses in the social sciences are never published or even written up, and the likelihood that a finding is shared with the broader research community falls sharply for “null” findings, i.e., those that are not statistically significant (Franco, Malhotra, and Simonovits 2014).

- Cleverly, the authors are able to look inside the file drawer through their access to the universe of studies that passed rigorous peer review for inclusion in a nationally representative social science survey administered at no cost to the researchers, namely, the National Science Foundation (NSF)-funded Time- sharing Experiments in the Social Sciences, or TESS.

- They find a striking empirical pattern: studies where the main hypothesis test yielded null results are 40 percentage points less likely to be published in a journal than a strongly statistically significant result, and a full 60 percentage points less likely to be written up in any form.

- Figure 1 reproduces some of the main patterns from Franco, Malhotra, and Simonovits (2014), as described in Mervis (2014b).

- Brodeur et al. (2016) collected a large sample of test statistics from papers in three top journals that publish largely empirical results (the American Economic Review, Quarterly Journal of Economics, and Journal of Political Economy) from 2005– 11.

- They propose a method to differentiate between the journals' selection of papers with statistically stronger results and inflation of significance levels by the authors themselves.

- Brodeur et al. (2016) document a rather disturbing two-humped density function of test statistics, with a relative dearth of reported *p*- values just above the standard 0.05 level (i.e., below a *t*- statistic of 1.96) cutoff for statistical significance, and greater density just below 0.05 (i.e., above 1.96 for *t*- statistics).

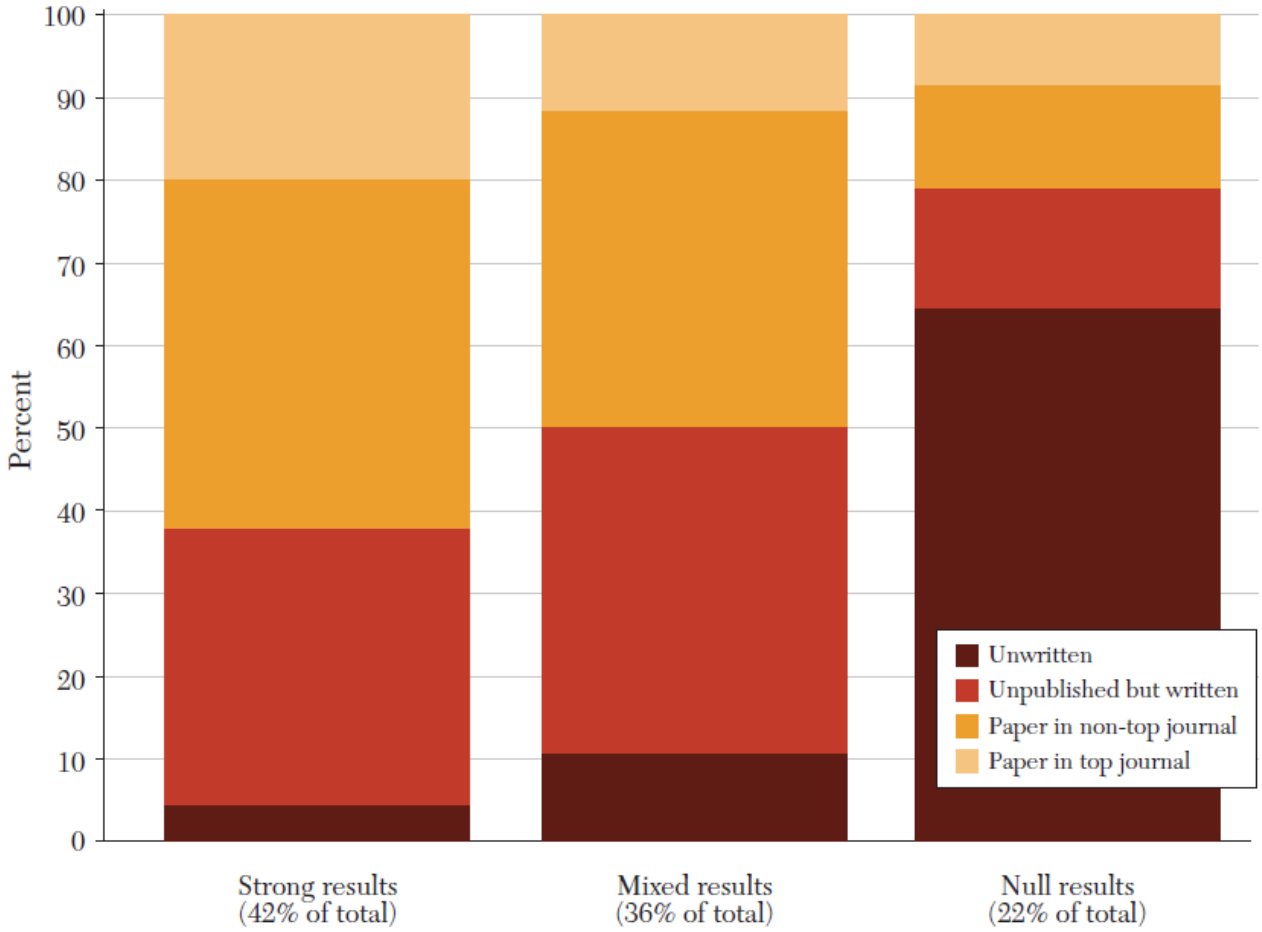

Figure 1. Publication Rates and Rates of Writing Up of Results from Experiments with Strong, Mixed, and Null Results

Source: Mervis (2014b). Reprinted with permission from AAAS. Experiments represent nearly the complete universe of studies conducted by the TESS.

2.3 Publication Bias and Effect Size

TABLE 2
EXAMPLES OF RECENT META-ANALYSES IN ECONOMICS

Paper	Topic	Publication bias?	Papers (Estimates) used	Notes
Brodeur et al. (2016)	Wide collection of top publications	+	641 (50,078)	Finds that 10–20% of significant results are misplaced, and should not be considered statistically significant.
Vivalt (2015)	Developing country impact evaluation	+	589 (26,170)	Finds publication bias/specification search is more prevalent in non-experimental work.
Viscusi (2015)	Value of a statistical life (VSL)	+	17 (550)	Use of better and more recent fatality data indicates publication bias exists, but that accepted VSL are correct.
Doucouliagos, Stanley, and Viscusi (2014)	VSL and income elasticity	+	14 (101)	Previous evidence was mixed, but controlling for publication bias shows the income elasticity of VSL is clearly inelastic.
Doucouliagos and Stanley (2013)	Meta-meta-analysis	+	87/3,599 (19,528)	87 meta analyses with 3,599 original articles and 19,528 estimates show that 60% of research areas feature substantial or severe publication bias.
Havranek and Irsova (2012)	Foreign direct investment spillovers	~	57 (3,626)	Find publication bias only in published papers and only in the estimates authors consider most important.
Mookerjee (2006)	Exports and economic growth	+	76 (95)	Relationship between exports and growth remains significant, but is significantly smaller when corrected for publication bias.
Nijkamp and Poot (2005)	Wage curve literature	+	17 (208)	Evidence of publication bias in the wage curve literature (the relationship between wages and local unemployment); adjusting for it gives an elasticity estimate of −0.07 instead of the previous consensus of −0.1.

TABLE 2
EXAMPLES OF RECENT META-ANALYSES IN ECONOMICS

Paper	Topic	Publication bias?	Papers (Estimates) used	Notes
Abreu, de Groot, and Florax (2005)	Growth rate convergence	0	48 (619)	Adjusting for publication bias in the growth literature on convergence does not change estimates significantly.
Doucouliagos (2005)	Economic freedom and economic growth	+	52 (148)	Literature is tainted, but relationship persists despite publication bias.
Rose and Stanley (2005)	Trade and currency unions	+	34 (754)	Relationship persists despite publication bias. Currency union increases trade 30–90%.
Longhi, Nijkamp, and Poot (2005)	Immigration and wages	0	18 (348)	Publication bias is not found to be a major factor. The negative effect of immigration is quite small (0.1%) and varies by country.
Knell and Stix (2005)	Income elasticity of money demand	0	50 (381)	Publication bias does not significantly affect the literature. Income elasticities for narrow money range from 0.4 to 0.5 for the United States and 1.0 to 1.3 for other countries.
Doucouliagos and Laroche (2003)	Union productivity effects	+	73 (73)	Publication bias is not considered a major issue. Negative productivity associations are found in the United Kingdom, with positive associations in the United States.
Gorg and Strobl (2001)	Multinational corporations and productivity spillovers	+	21 (25)	Study design affects results, with cross-sectional studies reporting higher coefficients than panel data studies. There is also some evidence of publication bias.
Ashenfelter, Harmon, and Oosterbeek (1999)	Returns to education	+	27 (96)	Publication bias is found, and controlling for it significantly reduces the differences between types of estimates of returns to education.

Notes: Table shows a sample of recent papers conducting meta-analyses and testing for publication bias in certain literatures in economics. Positive evidence for publication bias indicated by "+", no evidence for publication bias with "0", and mixed evidence with "~". The number of papers and total estimates used in the meta-analysis are also shown.

2.3.1 Subgroup Analysis

2.4 Inability to Replicate Results

2.4.1 Data Availability

TABLE 3
TRANSPARENCY POLICIES AT SELECTED TOP ECONOMICS AND FINANCE JOURNALS

Journal	Data-sharing policy?	Notes	Replication/ comment publication?	Notes	Funding/conflict of interest disclosure?	Notes
American Economic Review	Yes	Current policy was announced in 2004, becoming effective in 2005. It is in effect for all AEA journals.	Yes		Yes	Implemented in July 2012 for all AEA journals.
American Economic Journals (Applied Economics; Economic Policy; Macroeconomics)	Yes	Same as *AER*. Since journal inception in 2009.	Yes	Allow post-publication peer review on website.	Yes	Same as *AER*.
Econometrica	Yes	Began in 2004. See Dekel et al. (2006).	Yes		Yes	Peer review conflict of interest statement printed January 2009. Current financial disclosure policy adopted May 2014.
Journal of Finance	No		Yes		Yes	Current policy adopted August 2015.

TABLE 3
TRANSPARENCY POLICIES AT SELECTED TOP ECONOMICS AND FINANCE JOURNALS

Journal	Data-sharing policy?	Notes	Replication/comment publication?	Notes	Funding/conflict of interest disclosure?	Notes
Journal of Financial Economics	No	Some data is available on the journal webpage, but there appears to be no official policy.	No		Yes	Current policy adopted November 2015.
Journal of Political Economy	Yes	Uses the same policy as the *AER*. Announced in 2005, effective in 2006.	Yes	Submission instructions state that authors of comments must correspond with original authors.	No	
Quarterly Journal of Economics	Yes	Uses the same policy as the *AER*, adopted 2016.	Yes		Yes	
Review of Economic Studies	Yes	Start date unclear.	No		No	
Review of Financial Studies	No		Yes		Yes	Adopted August 2006. Updated June 2016.

Notes: These eleven journals are at the top of the Scientific Journal Rankings (SJR), excluding the *Journal of Economic Literature*, since its publications are generally reviews; see http://www.scimagojr.com/journalrank.php?area=2000. The *American Economic Journal: Microeconomics* has the same policies as the other *AEJ* journals, but is lower ranked. Data-sharing policy indicates whether the journal has a policy requiring authors to submit data that produces final results. Information obtained from journal websites and instructions for authors as well as via email to journal staff through October 2016. Replication/comment publication indicates whether the journal has published a replication, as per Duvendack, Palmer-Jones, and Reed (2015) or The Replication Network list (http://replicationnetwork.com/replication-studies/) as well as journal websites. Since "replication" is an imprecise term, this categorization is perhaps subject to some debate.

2.4.1.1 Proprietary Data

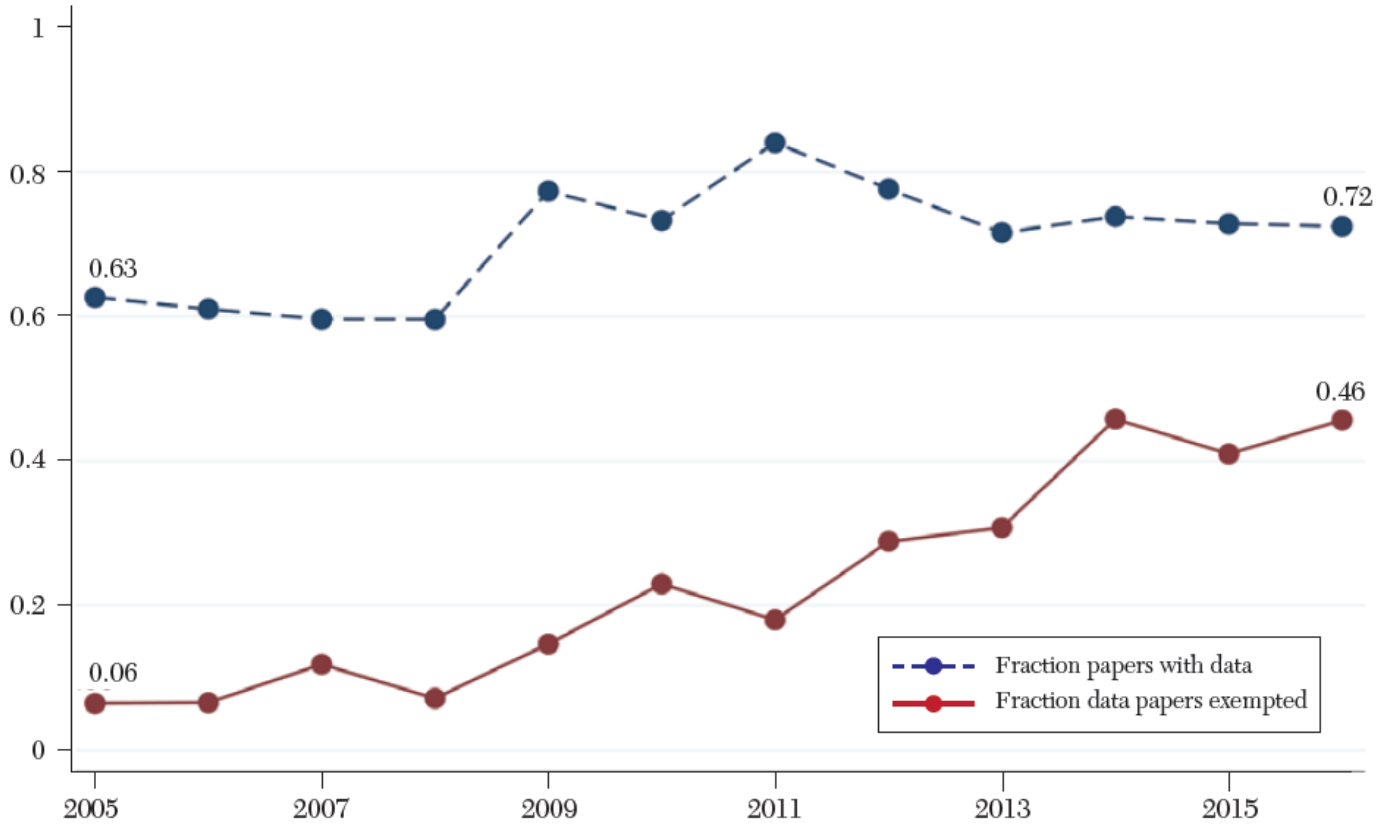

Figure 2. *AER* Papers with Data Exempt from the Data-Sharing Requirement

Note: Figure shows annual data on the fraction of *American Economic Review* papers that use data, and the fraction of those data-using papers that were exempted from the data-sharing policy.

Source: Data is taken from the Annual Report of the Editors, which appears annually in the *Papers and Proceedings* issue of the *AER*. Figure available in public domain: http://dx.doi.org/10.7910/DVN/FUO7FC.

2.4.2 Types of Replication Failures and Examples

2.4.2.1 Evidence on Replication in Economics

- One of the most important recent studies is Camerer et al. (2016), which repeated eighteen behavioral economics lab experiments originally published between 2011 and 2014 in the *American Economic Review* and the *Quarterly Journal of Economics* to assess their replicability.

- Figure 3 below reproduces a summary of their findings.

- Their approach is similar in design to a large-scale replication of one hundred studies in psychology known as the "Replication Project: Psychology," (RPP) which we discuss in detail below.

- In all, the estimated effects were statistically significant with the same sign in eleven of the eighteen replication studies (61.1 percent), albeit nearly always smaller in magnitude.

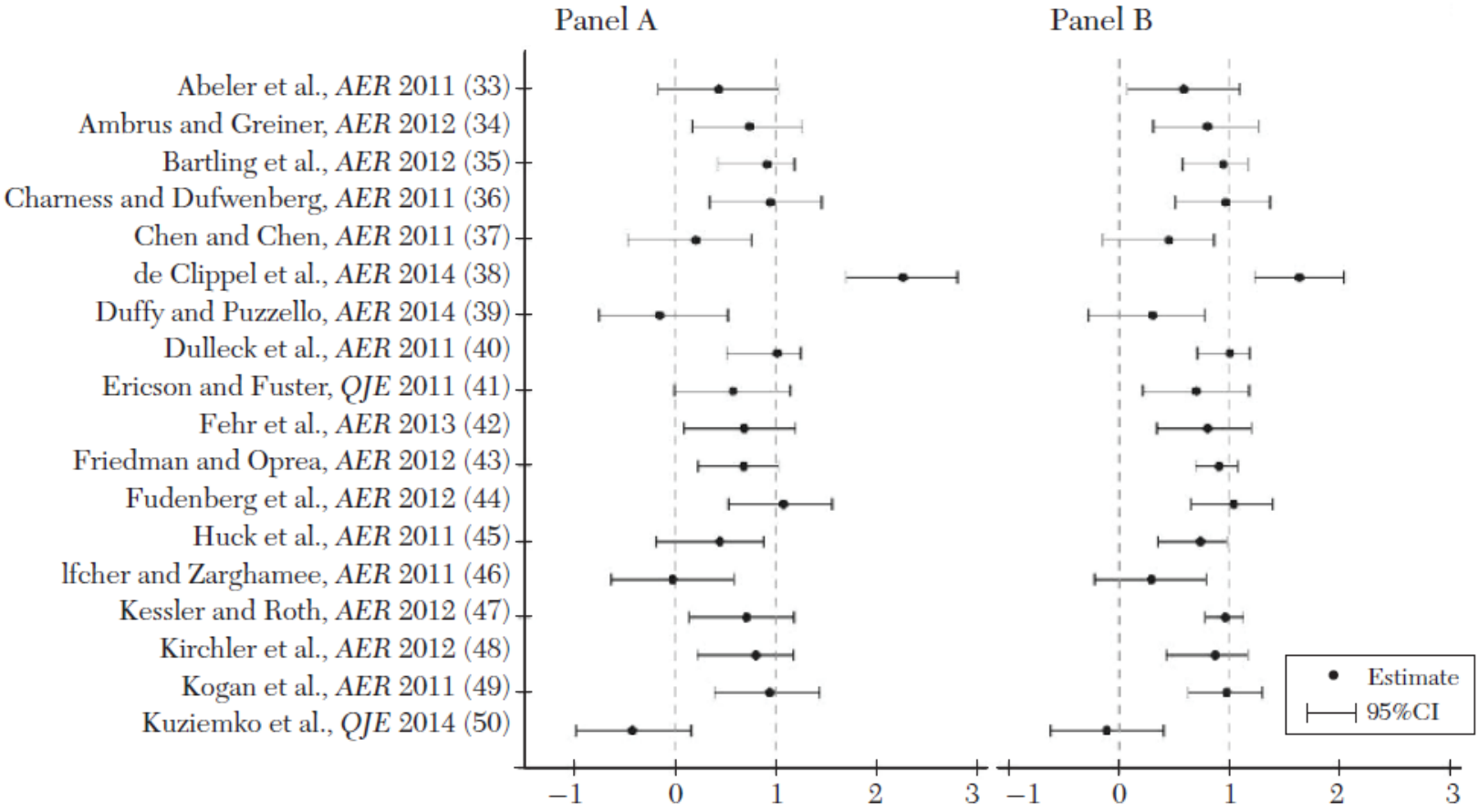

Figure 3. Replicability in Experimental Economics

2.4.2.2 Verification

2.4.2.3 Reproduction

2.4.2.4 Reanalysis

2.4.2.5 Extension

2.4.3 Fraud and Retractions

3. New Research Methods and Tools

3.1 Improved Analytical Methods: Research Designs and Meta- Analysis

3.1.1 Understanding Statistical Model Uncertainty

3.1.1.1 Model Averaging

$$(3) \qquad \hat{\delta}_M = \sum_m \mu(m \mid D)\, \hat{\delta}_m,$$

where m refers to a particular statistical model, M is the space of plausible models, $\mu(m \mid D)$ is the posterior probability of a model being the true model given the data D, and $\hat{\delta}_m$ is the estimated statistic from model $m \in M$.

3.1.1.2 The LSE School, Data Mining, and Machine Learning

$$(4) \qquad t_i = \frac{Estimated\ effect_i}{SE_i}$$

$$= \beta_0 + \beta_1 \left(\frac{1}{SE_i} \right) + v_i.$$

The resulting t-test on β_0, referred to as the Funnel Asymmetry Test (FAT) (Stanley 2008), captures the correlation between estimated effect size and precision, and thus tests for publication bias. This analysis

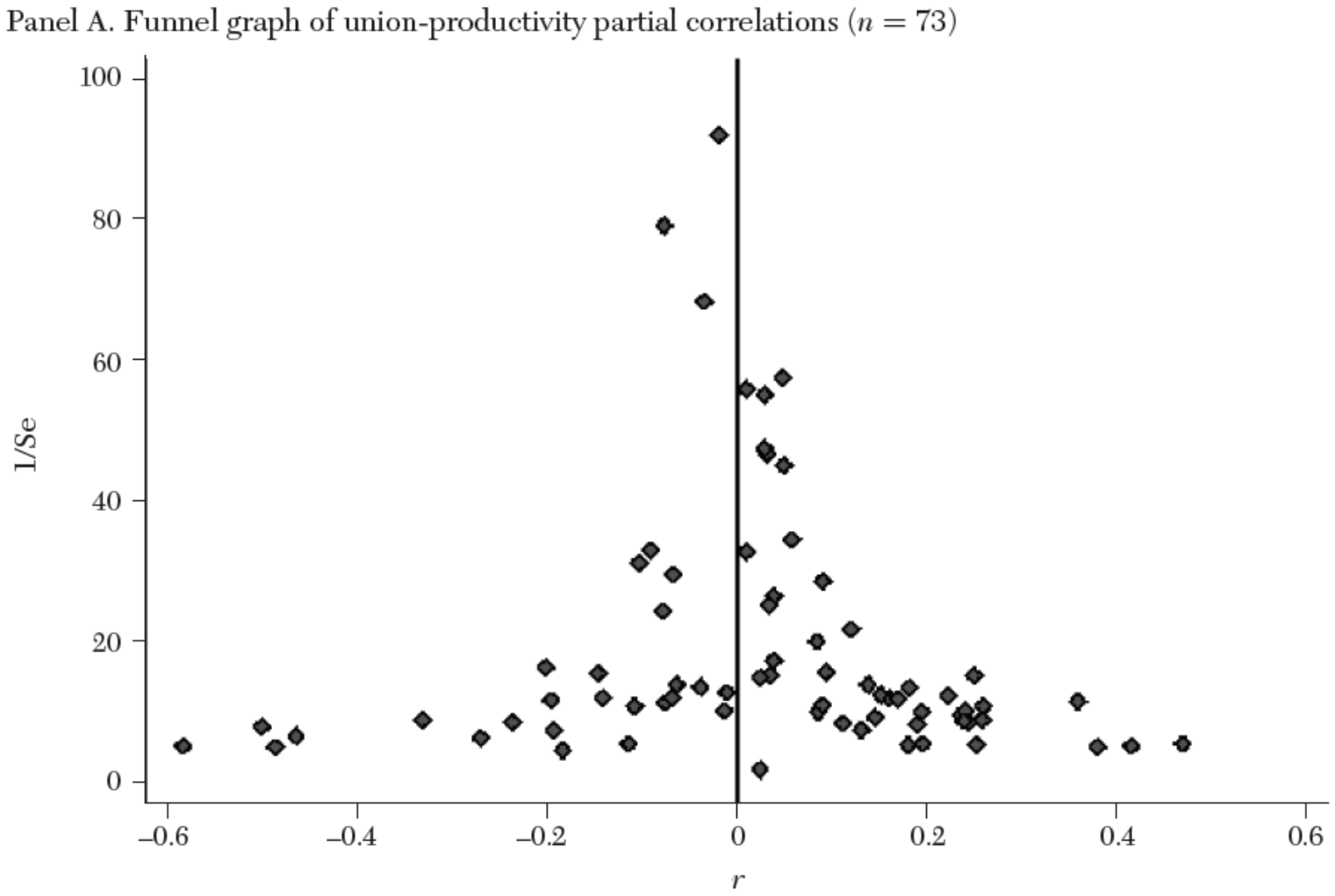

Figure 4. Examples of Funnel Graphs from the Union and Minimum-Wage Literature in Labor Economics

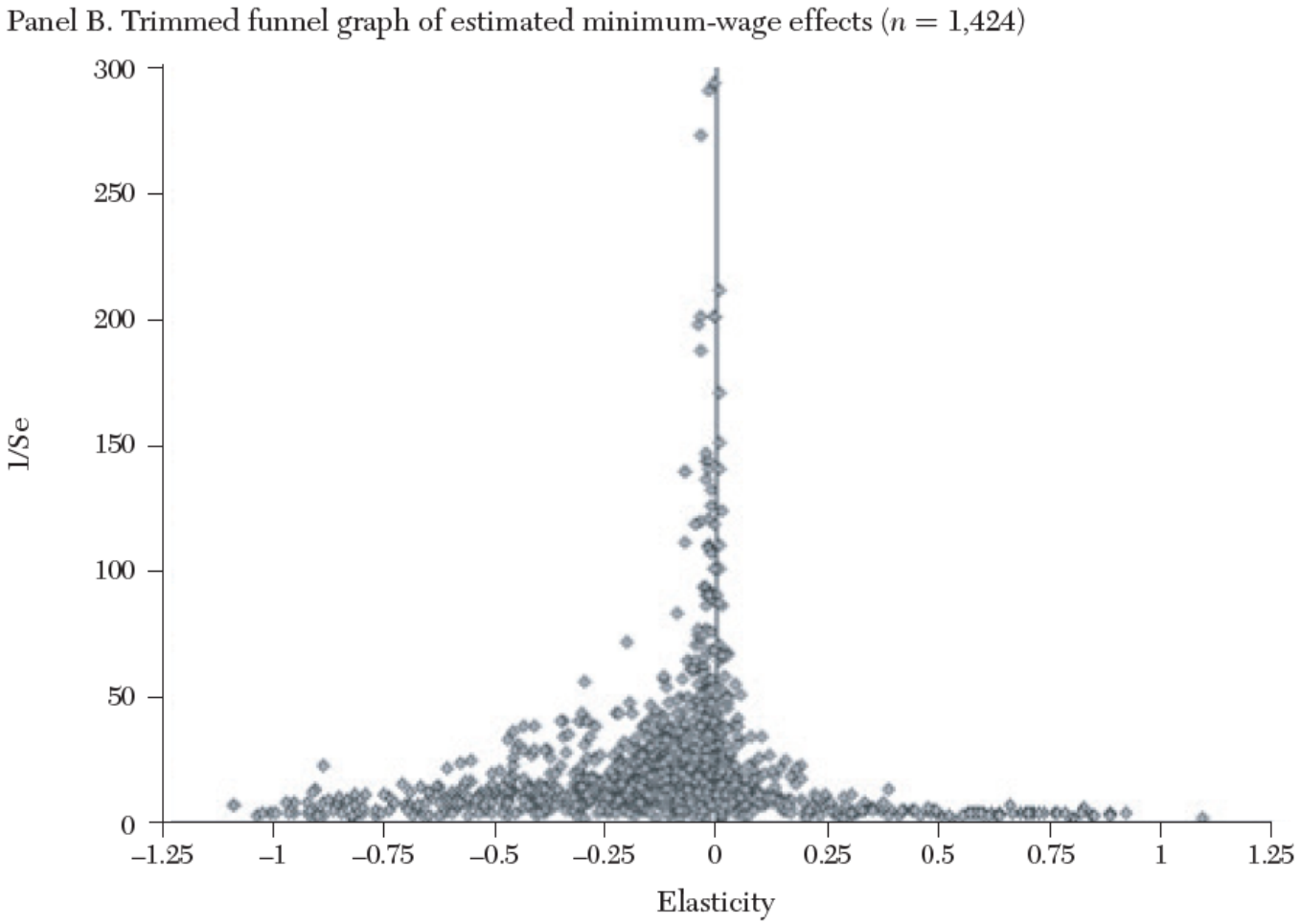

Figure 4. Examples of Funnel Graphs from the Union and Minimum-Wage Literature in Labor Economics

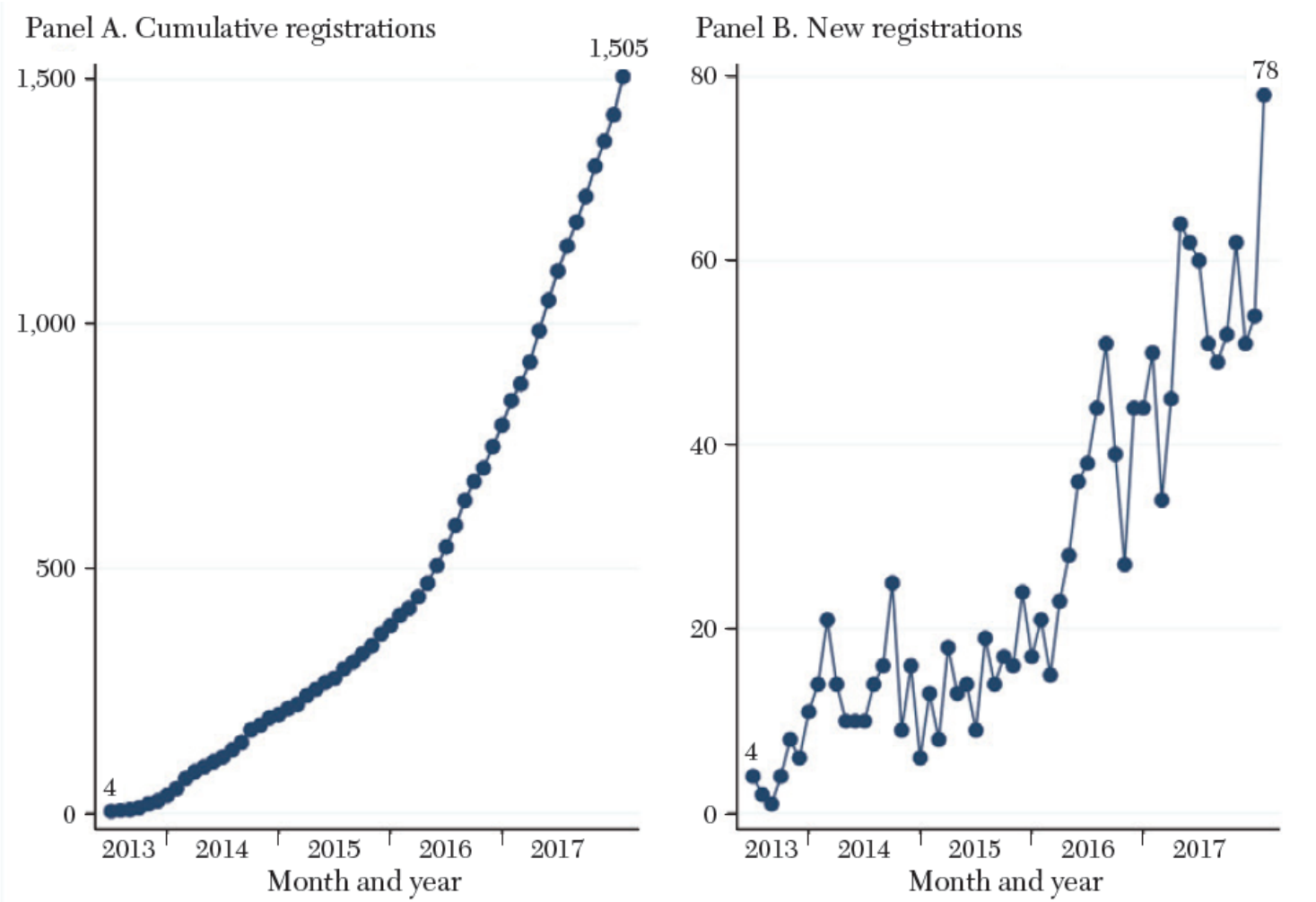

Figure 5. Studies in the AEA Trial Registry, May 2013 to December 2017

Notes: Figure shows the cumulative (panel A) and new (panel B) trial registrations in the American Economic Association Trial Registry (http://socialscienceregistry.org). Figure available in public domain: http://dx.doi.org/10.7910/DVN/FUO7FC.

TABLE 4
ERRONEOUS INTERPRETATIONS UNDER "CHERRY PICKING"

Outcome variable:	Mean in control group	Treatment effect	Standard error
Panel A. GoBifo "weakened institutions"			
Attended meeting to decide what to do with the tarp	0.81	−0.04+	(0.02)
Everybody had equal say in deciding how to use the tarp	0.51	−0.11+	(0.06)
Community used the tarp (verified by physical assessment)	0.90	−0.08+	(0.04)
Community can show research team the tarp	0.84	−0.12*	(0.05)
Respondent would like to be a member of the Village Development Committee	0.36	−0.04*	(0.02)
Respondent voted in the local government election (2008)	0.85	−0.04*	(0.02)
Panel B. GoBifo "strengthened institutions"			
Community teachers have been trained	0.47	0.12+	(0.07)
Respondent is a member of a women's group	0.24	0.06**	(0.02)
Someone took minutes at the most recent community meeting	0.30	0.14*	(0.06)
Building materials stored in a public place when not in use	0.13	0.25*	(0.10)
Chiefdom official did not have the most influence over tarpaulin use	0.54	0.06*	(0.03)
Respondent agrees with "Responsible young people can be good leaders"	0.76	0.04*	(0.02)
Correctly able to name the year of the next general elections	0.19	0.04*	(0.02)

Notes: Reproduced from Casey et al (2012, table VI). (i) Significance levels (per comparison p-value) indicated by $+p < 0.10$, $*p < 0.05$, $**p < 0.01$; (ii) robust standard errors; (iii) treatment effects estimated on follow-up data; and (iv) includes fixed effects for the district council wards (the unit of stratification) and the two balancing variables from the randomization (total households and distance to road) as controls.

4. Future Directions and Conclusions