Mass Reproducibility and Replicability: A New Hope

40th Annual Meeting of the European Economics Association 2025 Congress

Session on *Bias in Evaluation and Research Practices*

Monday August 25th, 2025

Derek Mikola

University of Ottawa

Institute for Replication

350+ Amazing Coauthors

Abel Brodeur, Derek Mikola, Nikolai Cook, Thomas Brailey, Ryan Briggs, Alexandra de Gendre, Yannick Dupraz, Lenka Fiala, Jacopo Gabani, Romain Gauriot, Joanne Haddad, Ryan McWay, Joel Levin, Magnus Johannesson, Lennard Metson, Jonas Minet Kinge, Wenjie Tian, Timo Wochner, Sumit Mishra, Joseph Richardson, Giulian Etingin-Frati, Alexi Gugushvili, Jakub Procházka, Myra Mohnen, Jakob Möller, Rosalie Montambeault, Sébastien Montpetit, Jason Collins, Sigmund Ellingsrud, Alexander Kustov, Louis-Philippe Morin, Todd Morris, Erlend Fleisje, Elaheh Fatemi-Pour, Scott Moser, Matt Woerman, Tim Ölkers, Edward Miguel, Fabio Motoki, Anders Kjelsrud, Lucija Muehlenbachs, Andreea Musulan, Christian Czymara, Hooman Habibnia, Alexander Coppock, Idil Tanrisever, Marco Musumeci, Nicholas Rivers, Miquel Oliver i Vert, Emre Oral, Alejandro Abarca, Christian Oswald, Ali Ousman, Marcin Wroński, Sonja Häffner, Ömer Özak, Filip-Mihai Toma, Myriam Brown, Diego Marino Fages, Shubham Pandey, Joseph Taoyi Wang, Andrea Erhart, Alexandre Pavlov, Bruno Rodrigues, William Roelofs, Jonathan D. Hall, Tobias Roemer, Ole Rogeberg, Julian Rose, Nadjim Fréchet, Maddalena Totarelli, Andrew Roskos-Ewoldsen, Shiang-Hung Hu, Alexandre Fortier-Chouinard, Paul Rosmer, Thomas Galipeau, Patrick Nüß, Ahwaz Akhtar, Andreas Kotsadam, Barbara Sabada, Goncalo Lima, Martín Brun, Soodeh Saberian, Van-Anh Tran, Nicolas Salamanca, Georg Sator, Andaleeb Rahman, Olle Hammar, Daniel Scates, Myra Yazbeck, Ben Couillard, Elmar Schlüter, Cameron Sells, Stephan Bruns, Sharmi Sen, Jori Korpershoek, Karim Nchare, Ritika Sethi, Kangyu Qiu, Eduardo Alberto Ramirez Lizardi, Alexa Federice, Yu-Shiuan Huang, Anna Shcherbiak, Moyosore Sogaolu, Matt Soosalu, Nino Buliskeria, Jonathan Créchet, Gustav Chung Yang, Erik Ø. Sørensen, Amund Hanson Kordt, Marie Connolly, Manali Sovani, Noah Spencer, Stefan Staubli, Lewis Krashinsky, Mathias Huebener, Renske Stans, Linh Phan, Anya Stewart, Felix Stips, Guidon Fenig, Jan Feld, Lorenzo Crippa, Luther Yap, Sabina Albrecht, Romeo Penheiro, Kieran Stockley, Stephenson Strobel, Ethan Struby, Andrea Calef, Christoph Huber, Endre Borbáth, John Tang, Hung Truong, Nikita Tsoy, Mojtaba Firouzjaeiangalougah, Kerem Tuzcuoglu, Diego Ubfal, Zubaria Andlib, Haley Daarstad, Laura Villalobos, Raphaël Jananji, Michael Jetter, Ali Elminejad, Nurlan Lalayev, Jill Caviglia-Harris, Julian Walterskirchen

Replication and Reproduction Matter

- **Replication is a Fundamental Pillar of Science**
  - Conclusions should be relatively consistent (repeated and predictable)
  - Trust in science is gained from replicability

- **Reproduction is Easier than Replication**
  - “Scientific Auditing”
  - If no reproduction, then why attempt to replicate?

Findings Presented Today

- ***Institute for Replication (I4R)*** **facilitates mass reproductions and replications**

- **About 85% of our sample is (computationally) reproducible**
- **26 (of 110) papers contain some form of coding error/discrepancy**
- **About 70% of re-analyses *remain* statistically significant at 5% and same sign**
  - **Heterogeneous rates by *type* of reanalysis**

- **Many Analysts Approach**
  - **Decreasing reproducibility as replicators experience increases**
  - **No relationship between data & code provisions and reproducibility**

Data for Main Results

- **110 reports with robustness reproductions or replications:**
  - **Very selected sample** (next slide)

- **About 5,000 new point estimates from the following re-analyses:**
  - (i) alternative choice of control variables
  - (ii) changing the sample
  - (iii) changing the dependent variable
  - (iv) changing the main independent variable
  - (v) changing the estimation method/model
  - (vi) changing the method of inference
  - (vii) change weighting scheme
  - (viii) replication using new data

Notes About Sample

- Targeted journals (e.g. data policies are enforced)
- Accessible data and codes
- Teams (“Replicators”) select papers which interest them
- Omit test statistics or coefficients which aren’t comparable

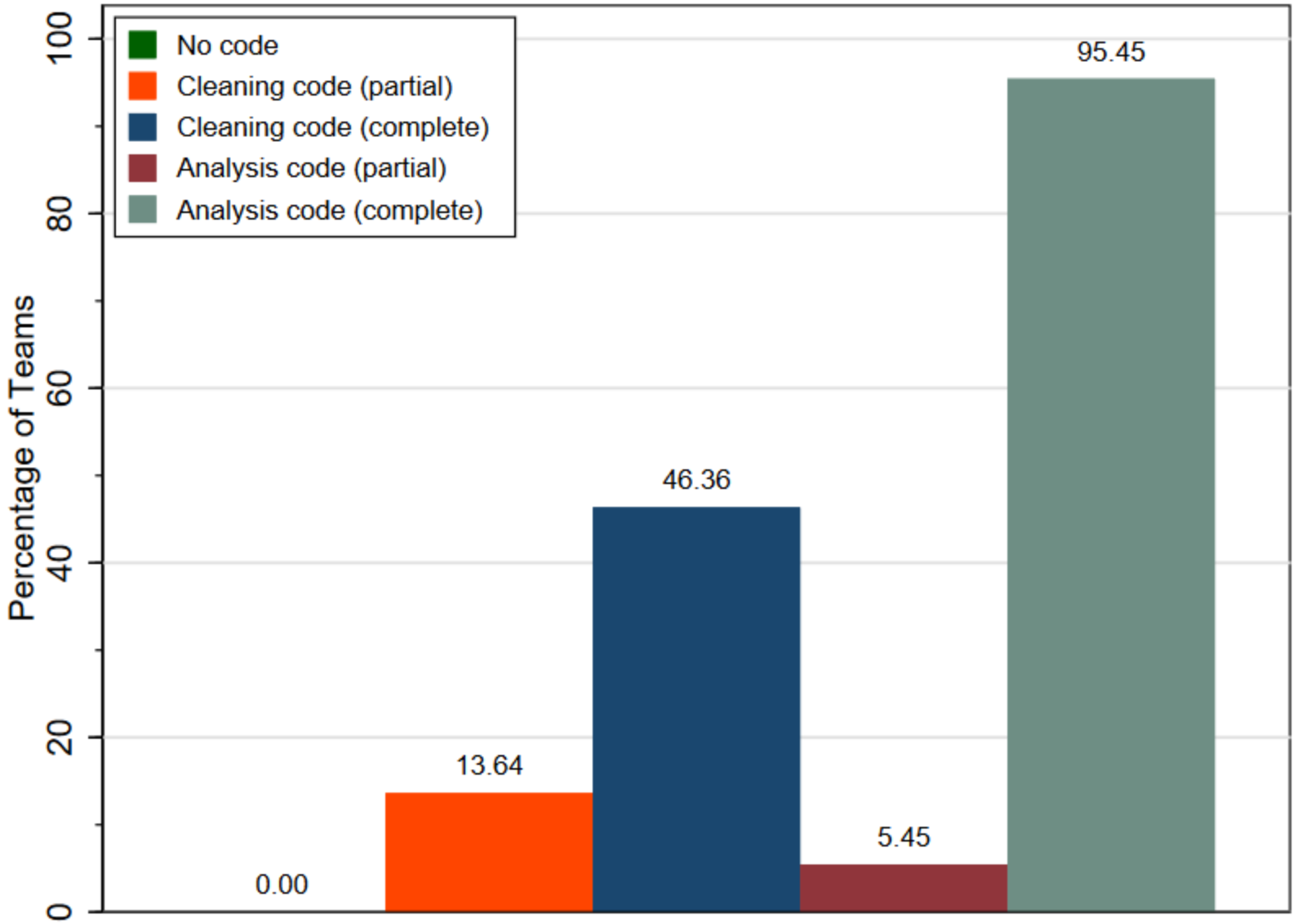
- Only studies since 2022
- Look primarily at main claims from a paper
- We somewhat umpire “reasonable” robustness checks
- Potential for revision after interacting with original authors

Summary Statistics

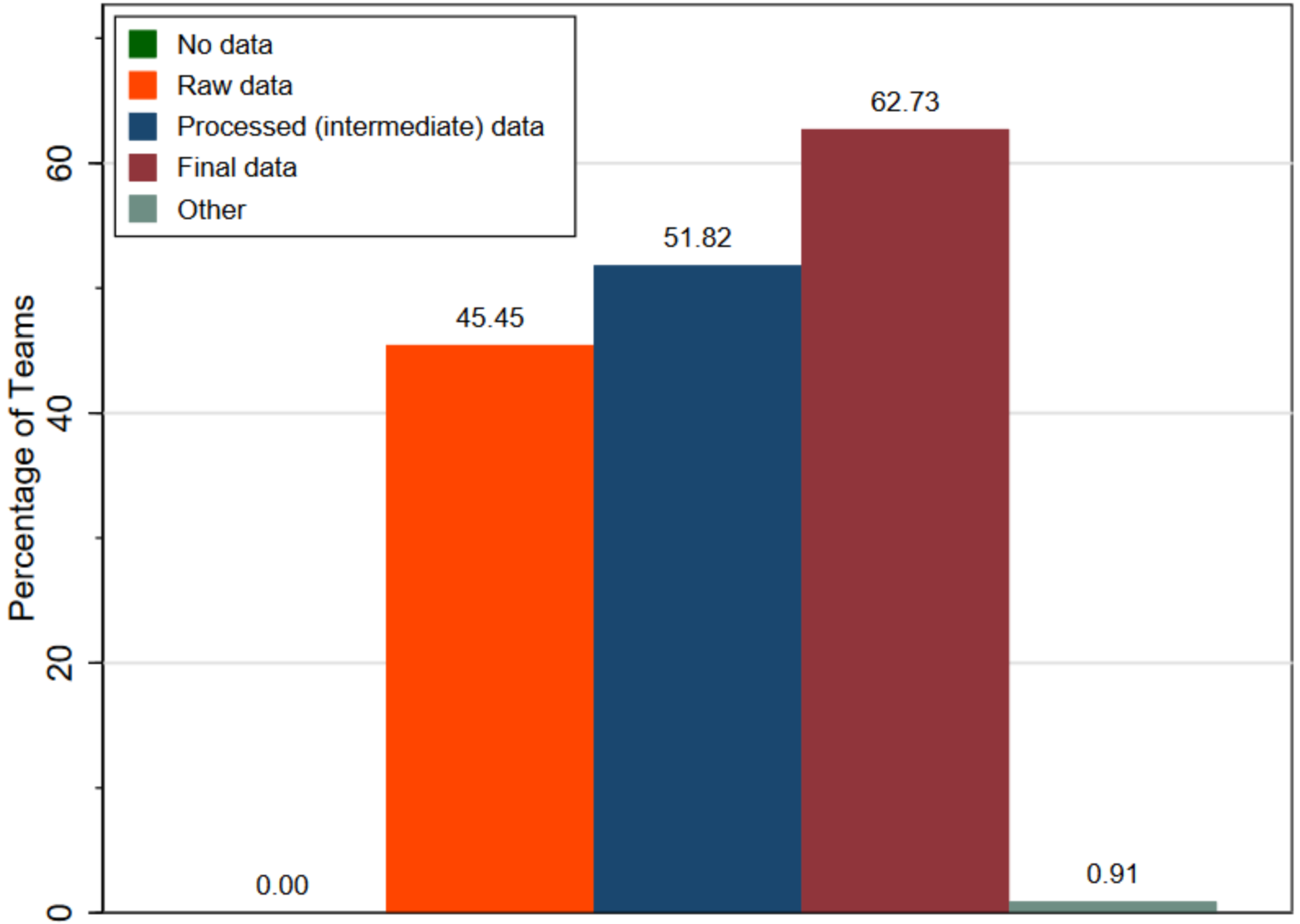
Table 1: Summary Statistics by Journal

Discipline and Journal	# Articles Total (1)	# Articles RGs (2)	# Articles Editor (3)	# Tests (4)	Data Editor (5)
Economics	79	67	12	5,494	
American Economic Review	17	12	5	1,392	Yes
American Economic Review: Insights	2	0	2	149	Yes
American Economic J.: Applied Economics	9	6	3	260	Yes
American Economic J.: Economic Policy	11	11	0	811	Yes
American Economic J.: Macroeconomics	3	3	0	25	Yes
Economic Journal	20	18	2	1,262	Yes
Journal of Political Economy	8	8	0	1,283	No
Quarterly Journal of Economics	4	4	0	101	No
Review of Economic Studies	5	5	0	211	Yes
Political Science	31	16	15	1,089	
American Journal of Political Science	13	6	7	539	External
American Political Science Review	6	3	3	214	No
Journal of Politics	12	7	5	336	Yes

Code Availability



Data Availability



First Meta Paper: Coding Errors and Deviations

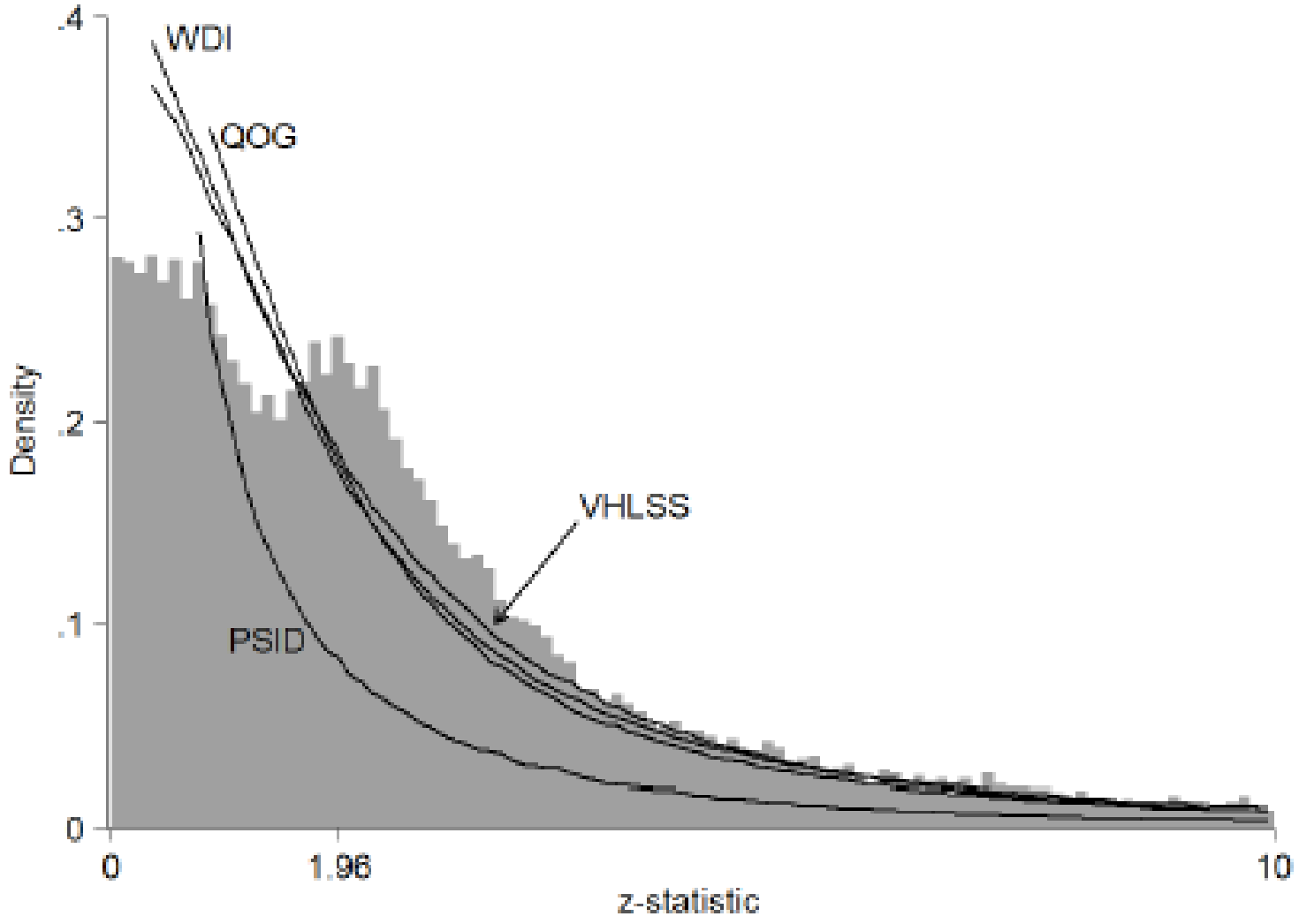
- **25 (of 110) papers have a coding error or deviation:**
  - Range from minor to major
    - » Ex. 75% of observations are duplicates
    - » Not cleaning raw data (e.g., St. Louis, St Louis, StLouis, …)
    - » Not fully interacting coefficients in a DID model
    - » Not specifying GMM function

- **Deviation: mentioning something in the paper, but doing something else in the code**
  - Harder to spot; less objective

- **Important coding decisions buried in footnotes or appendices**
- **Does not include *transcription* errors (rounding, fat-finger error)**

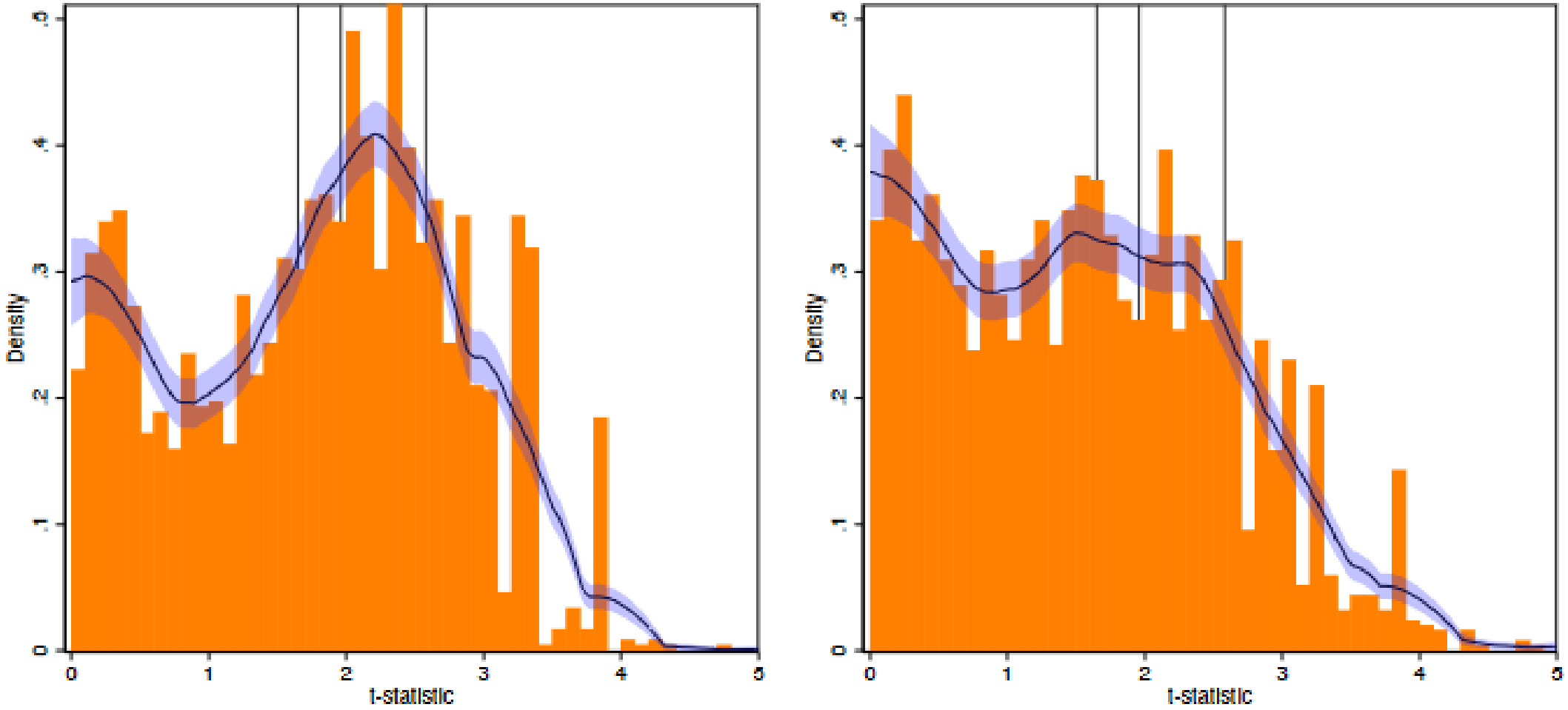
Recall: With and without P-Hacking

- Brodeur et al., 2016: Star Wars: Empirics Strike Back (AEJ:AE)
- Observed and counterfactual distributions



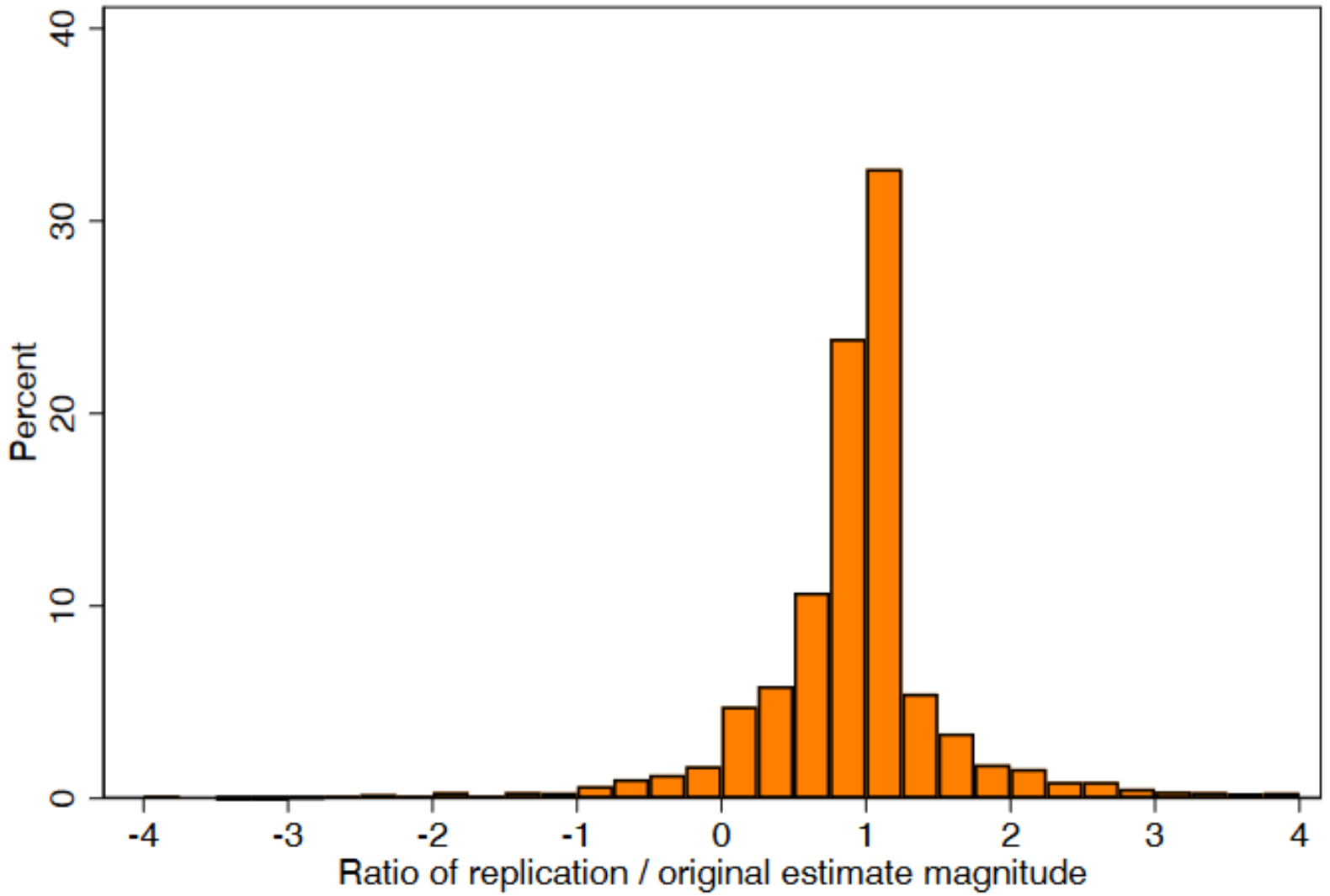
First Meta Paper: t-curves

Figure 3: Distributions of t-Statistics for Original Studies and Re-Analyses



First Meta Paper: Relative Effect Size

Figure 5: Relative Effect Size



Robustness Reproducibility Rate

Table below reads like a transition matrix

Table 4: Shifts in Statistical Significance Regions

	Re-Analysis Significance Level					
Original Significance Level	Sign Change	Not Sig.	Sig. at 10%	Sig. at 5%	Sig. at 1%	Total
Not Significant	13.61	75.00	4.59	3.91	2.89	100.00
Significant at 10%	6.91	45.45	28.00	12.73	6.91	100.00
Significant at 5%	2.76	27.89	12.06	41.08	16.21	100.00
Significant at 1%	4.95	12.89	4.43	8.07	69.66	100.00
Total	7.32	37.72	7.80	14.06	33.10	100.00

For Original Results which were statistically significant at the 5% level,

About 70% of reanalysis remain statistically significant at the 10% level

Robustness Reproducibility Rate

- **Barriers to sensitivity analysis:**
  - Self-report: by far the main barrier is the lack of raw data

- **Re-analyses by Type (Appendix Table)**
  - **Lowest robustness reproducibility rates:** changing the dependent variable, sample
  - **Highest:** using new data
  - **Middle range:** changing controls, estimation method, inference method, main independent variable, weighting scheme

Many Analysts Approach

- Asked teams of researchers to evaluate prepared questions with prepared datasets.

- Most teams find a negative relationship between replicators' experience and reproducibility

- No relationship between reproducibility and the provision of intermediate or even raw data combined with the necessary cleaning codes

Conclusion

- **Do we have a reproducibility and replicability crisis?**
  - **Yes:** Lots of p-hacking and coding errors. And this is only marginal p-hacking and researchers don't even share (all) their data. Probably worse for lower ranked journals and authors who don't share
  - **No:** We don't see a lot of p-hacking in those figures. Reproducibility rates look high, which is reassuring. We (econ/pol sci) are doing much better than other disciplines, which is probably due to our training
- **Mass reproducibility may have a positive impact on views of the discipline:**
  - 40% of replicators report that the quality of the replication package led them to have a more optimistic view of the discipline
  - Another 40% reported no impact on their views

Appendix

Summary Statistics

Table 2: Summary Statistics: Original Authors and Replicators

	Mean (1)	Standard Deviation (2)	Minimum (3)	Maximum (4)
Test Statistics per Report	59.84	72.67	0	421
Year	2022.13	0.33	2022	2023
Economic Articles	0.72	0.45	0	1
Proportion of Economics Papers in Top 5	0.43	0.50	0	1
GS Citations (As of Report Completed)	43.98	71.39	0	573
Original Authors				
Number Original Authors	2.63	1.23	1	6
Share Graduate Student	0.06	0.18	0	1
Avg. Experience (Years since PhD)	11.21	6.34	0	31.50
Avg. GS Citations	4269.05	8882.00	31	55633.5
Replicators				
Number Replicators	3.25	1.22	1	7
Share Published Top 5 Econ/Targeted Poli Sci	0.15	0.36	0	1
Share Pub. Targeted Journals	0.30	0.46	0	1
Share Pub. Top 5/Targeted Poli Sci (Past 5 Years)	0.14	0.34	0	1
Share Pub. Targeted Journals (Past 5 Years)	0.26	0.44	0	1
Share Team Graduate Student	0.49	0.34	0	1
Avg. Experience (Years since PhD)	3.12	3.10	0	13.50
Avg. GS Citations	478.49	1016.67	0	6095.33
Comfortable programming in Stata	0.74	0.44	0	1
Comfortable programming in R	0.64	0.48	0	1
Comfortable programming in MATLAB	0.14	0.34	0	1

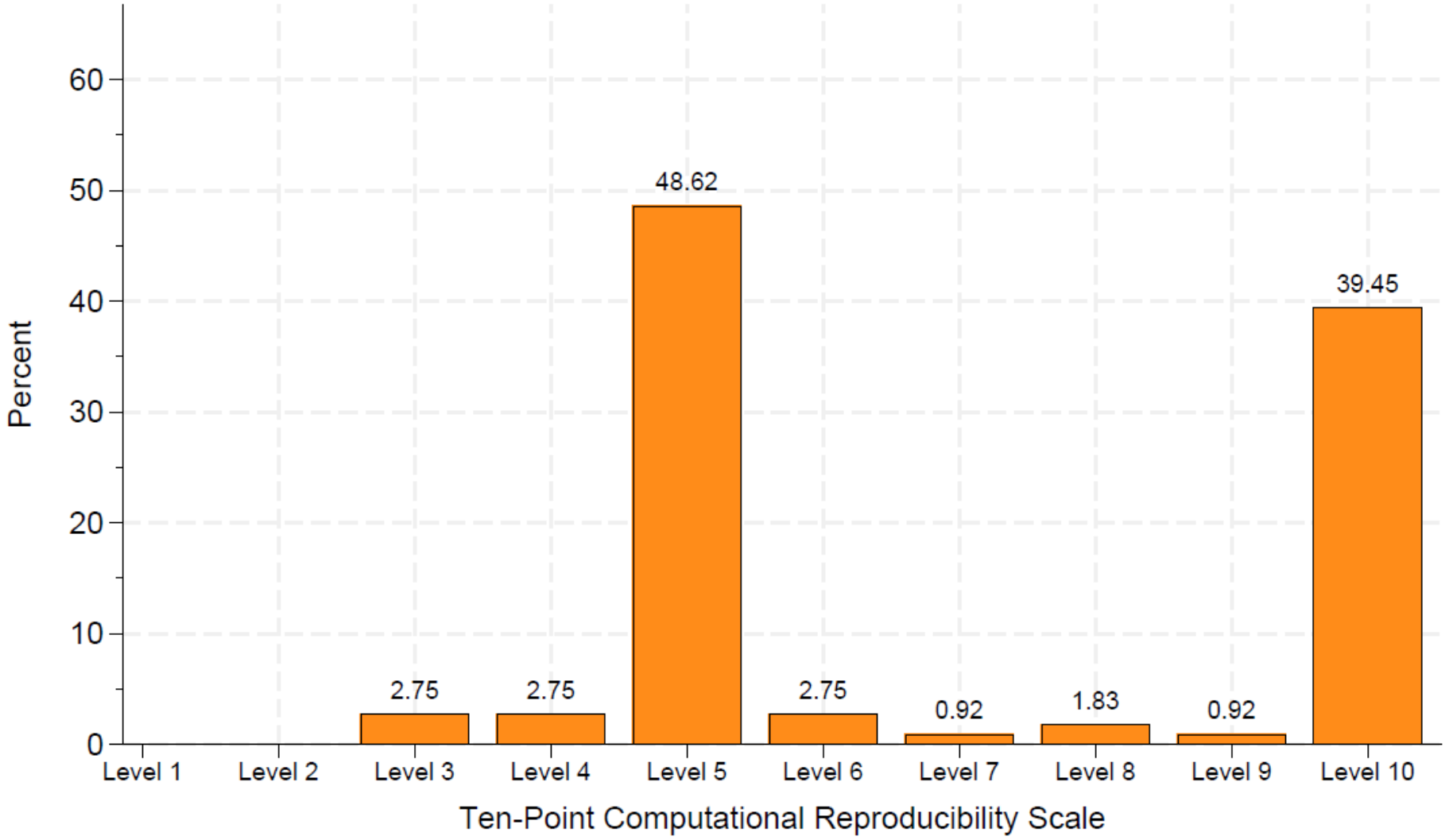
Summary Statistics

Table 3: Summary Statistics by Types of Re-Analyses

	# Articles (1)	# Articles RGs (2)	# Articles Editor (3)	# Tests (4)
All Re-Analyses	103	81	22	6583
All Simultaneous Robustness Checks	51	41	10	809
Full Sample				
By Re-Analyses: Change in				
Control variables	58	45	13	1939
Sample	75	57	18	1774
Dependent Variable	23	18	5	285
Main Independent Variable	20	19	1	264
Estimation Method	33	28	5	605
Inference Method	23	19	4	542
Weighting Scheme	14	10	4	126
Use New Data	15	13	2	469

Computational Reproducibility

Figure 1: 10-Point Computationally Reproducibility Score



Reproducibility by Type of Robustness Check

Table 13: Robustness Reproducibility and Replicability Rates (with counts)

	(1) Full Sample	(2) Change Control	(3) Dep. Var.	(4) Change Estim.	(5) Infer. Method	(6) Ind. Var.	(7) Change Sample	(8) Change Weights	(9) New Data
Rep. if Orig. Sig. 5%									
Estimates	0.71	0.76	0.45	0.76	0.74	0.78	0.64	0.74	0.87
Confidence Intervals	[0.70,0.73]	[0.73,0.79]	[0.35,0.55]	[0.72,0.81]	[0.67,0.82]	[0.72,0.85]	[0.61,0.67]	[0.64,0.85]	[0.84,0.91]
Observations	2552	833	96	348	121	160	945	66	370
Rep. if Orig. Not Sig. 5%									
Estimates	0.88	0.92	0.80	0.85	0.88	0.77	0.86	0.97	0.83
	[0.87,0.90]	[0.89,0.94]	[0.64,0.96]	[0.80,0.90]	[0.83,0.94]	[0.64,0.89]	[0.83,0.89]	[0.91,1.03]	[0.75,0.91]
Observations	1453	594	25	174	129	47	468	33	83
Rep. if Orig. Sig. 10%									
Estimates	0.75	0.78	0.45	0.83	0.74	0.80	0.70	0.73	0.89
Confidence Intervals	[0.74,0.77]	[0.75,0.81]	[0.36,0.55]	[0.79,0.86]	[0.67,0.82]	[0.74,0.86]	[0.67,0.73]	[0.63,0.83]	[0.86,0.92]
Observations	2826	932	106	373	137	168	1068	74	382
Rep. if Orig. Not Sig. 10%									
Estimates	0.85	0.88	0.93	0.82	0.84	0.54	0.82	0.92	0.75
Confidence Intervals	[0.83,0.87]	[0.85,0.91]	[0.80,1.06]	[0.76,0.88]	[0.77,0.91]	[0.38,0.70]	[0.78,0.86]	[0.81,1.03]	[0.64,0.85]
Observations	1179	495	15	149	113	39	345	25	71