p-Hacking and the Replication Crisis

With an Emphasis on Economic Applications

Prepared for Hedibert Lopes
Professor of Statistics and Econometrics
Insper Institute of Education and Research[1]

April 29, 2026

[1] Claude for Mac, version 1.4758.0 (fb266c)

Outline

What is p-Hacking?

p-Hacking (also: data dredging, specification searching) is the practice of manipulating data analysis until non-significant results become statistically significant, typically below the conventional $p < 0.05$ threshold.

Formal idea

Under the null H_0, p-values should be uniform on $[0, 1]$. Selectively searching across models, samples, or variables *distorts* this distribution, producing a discontinuity at $p = 0.05$ (Brodeur et al., 2016).

Key insight: a single significant result, mined from many silent attempts, is no longer evidence against H_0.

Common p-Hacking Practices

- **Specification searching:** trying many regressions, reporting only the "cleanest."
- **Selective control variables:** adding/removing covariates to push p below 0.05.
- **Sample slicing:** dropping outliers, choosing time windows, restricting subsamples.
- **Outcome switching:** reporting whichever dependent variable "works."
- **Optional stopping:** collecting data until significance appears.
- **HARKing:** Hypothesizing After the Results are Known.

The Replication Crisis

A growing body of evidence shows that many published findings fail to replicate in independent samples.

- Open Science Collaboration (2015): only ∼36% of 100 psychology studies replicated.
- Camerer et al. (2016, 2018): replication rates of 61% in experimental economics and 62% in social science *Nature*/*Science* papers.
- Effect sizes in successful replications are typically ∼50% of original estimates.

p-Hacking is widely viewed as a leading driver, alongside publication bias, low statistical power, and weak pre-registration norms.

Why p-Hacking Fuels the Crisis

1. Journals reward novel, "significant" findings $\Rightarrow$ publication bias.
2. Researchers face strong incentives to find $p < 0.05$.
3. Flexibility in the "garden of forking paths" (Gelman & Loken, 2014) makes Type I error rates far higher than nominal 5%.
4. The published literature becomes systematically biased upward in effect size and downward in standard errors.
5. Subsequent studies, meta-analyses, and policy decisions inherit the bias.

Evidence of p-Hacking in Economics

Brodeur, Lé, Sánchez, & Cook (2016, AEJ:Applied) examined 50,000+ test statistics from top economics journals (*AER*, *QJE*, *JPE*):

- A pronounced *two-humped* distribution of z-statistics around the 1.96 critical value.
- Suggests roughly 10–20% of marginally significant results are explained by p-hacking and selective reporting.

Brodeur, Cook, & Heyes (2020, AER) compared methods:

- Instrumental Variables (IV) and Difference-in-Differences (DiD) show the *most* evidence of inflation near $p = 0.05$.
- RCTs and Regression Discontinuity Designs (RDD) show *less*.

Why Economics is Particularly Vulnerable

- **Observational data & flexible identification:** many plausible specifications, instruments, and controls.
- **Limited replication culture:** until recently, replications were rarely published or rewarded.
- **High-stakes "stars" ($*$, $**$, $***$):** significance thresholds function as gatekeepers to publication.
- **Long lags between data collection and publication**, weakening pre-specification of hypotheses.
- **Policy relevance amplifies the cost** of false positives in labor, development, finance, and macro.

Notable Economic Examples

- **Minimum-wage employment effects:** highly sensitive to specification and sample choice (Neumark & Wascher vs. Card & Krueger).
- **Growth regressions:** Sala-i-Martin (1997) ran 2 million regressions to show how fragile cross-country growth determinants are.
- **Reinhart & Rogoff (2010):** "90% debt-to-GDP threshold" result reversed after data and coding errors uncovered (Herndon et al., 2014).
- **Behavioral / experimental economics:** Many Labs and replication projects (Camerer et al.) reduced confidence in several headline effects.

Responses in the Economics Profession

- **Pre-registration** and pre-analysis plans (esp. for RCTs via the AEA RCT Registry, since 2013).
- **Mandatory data and code archives** at *AER*, *QJE*, *Econometrica*, *ReStat*, etc.
- **Replication journals and sections** (e.g., *Journal of Comments and Replications in Economics*).
- **Multiple-hypothesis corrections** (Bonferroni, Romano-Wolf, sharpened FDR by Anderson, 2008).
- **Robustness reporting:** specification curves (Simonsohn et al.), multiverse analyses, and "many analysts" studies.
- **Lower / contextual significance thresholds** (Benjamin et al., 2018: redefine "significant" as $p < 0.005$).

Takeaways

- p-Hacking arises from *flexibility* in analysis combined with *strong incentives* to obtain $p < 0.05$.
- It is a key contributor to the broader replication crisis across the empirical sciences.
- In economics, evidence of p-hacking is concentrated in observational designs (IV, DiD); experimental and RDD studies fare better.
- Pre-registration, transparent reporting, replication, and stricter thresholds are the main lines of defense.
- The fundamental fix is cultural: rewarding *credible* research, not just *significant* research.

Selected References I

- Brodeur, A., Lé, M., Sánchez, M., & Cook, J. (2016). Star Wars: The Empirics Strike Back. *AEJ: Applied Economics*, 8(1), 1–32.
- Brodeur, A., Cook, N., & Heyes, A. (2020). Methods Matter: *p*-Hacking and Publication Bias in Causal Analysis in Economics. *AER*, 110(11), 3634–60.
- Camerer, C. et al. (2016). Evaluating Replicability of Laboratory Experiments in Economics. *Science*, 351(6280), 1433–1436.
- Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist*, 102(6), 460.
- Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science*, 349(6251), aac4716.
- Herndon, T., Ash, M., & Pollin, R. (2014). Does High Public Debt Consistently Stifle Economic Growth? *Cambridge J. of Economics*, 38(2).
- Benjamin, D. et al. (2018). Redefine Statistical Significance. *Nature Human Behaviour*, 2, 6–10.