Uncovering P-Hacking and Publication Bias in Causal Analysis: Insights from Top 25 Economic Journals*

P-Hacking Variation Across Method, Year, and Journal Rank: A Reproduction Study of American Economic Review

Chenhang Huang Zixuan Yang

February 15, 2024

This study delves into the integrity of empirical economic research by investigating the prevalence of p-hacking, a practice where researchers manipulate statistical tests to achieve desired results. Analyzing data from 25 economic journals between 2015 and 2018, we scrutinized the distribution of statistical test results across various methods and journal rankings. Our findings reveal a declining trend in p-hacking over time, shedding light on potential improvements in research practices. This study underscores the importance of transparency and rigor in empirical research, ensuring the credibility and reliability of scientific findings in economics and beyond.

1 Introduction

In empirical economics, establishing credible causal effect is paramount. ***Causal Effect Estimand: The average treatment effect (ATE) of a Treatment on outcomes.*** Randomized controlled trials (RCTs) provide a rigorous method for causal inference through experimental approaches. However, RCTs often face challenges such as high costs and ethical constraints, which may render them impractical.As alternatives, quasi-experimental methods like Difference-in-Differences (DID), Instrumental Variables (IV), and Regression Discontinuity Design (RDD) have been proposed for causal inference. Yet, the validity of these methods relies on specific assumptions, necessitating careful data handling by economists to meet these

*Code and data are available at: https://github.com/chenhanghuang/replication-research/tree/main/PHacking. Replication aspects are available at: https://www.socialsciencereproduction.org/reproductions/1635/select-paper

requirements. The flexibility in data analysis poses a risk of "p-hacking", where researchers selectively analyze data or conduct statistical tests to achieve desired levels of significance (typically a p-value below 0.05). When authors fail to adequately justify their data analysis methods, it can lead to false-positive results and undermine the credibility of scientific research.Authors may be inclined to seek significant results due to the perceived association between significance and publication acceptance. They may believe that significant findings are more likely to be accepted and published, thus finding significance becomes attractive.

Brodeur et al. (2016) used 50,000 tests published in the AER, JPE, and QJE between 2005 and 2011 and found that the distribution of p-values exhibited a camel shape. A large proportion of p-values were higher than 0.25, with a trough between 0.25 and 0.10, and a peak below 0.05. This phenomenon suggests that researchers may attempt to inflate the value of their tests by selectively reporting the most statistically significant results.However, whether these patterns persist in economic journals after 2011 needs further analysis, and this study aims to address this gap.

In this study, we aim to investigate the relationship between inference methods and statistical significance. We collected data by Brodeur, Cook, and Heyes (2020) from 25 economic journals, encompassing all articles published between 2015 and 2018. We recorded information such as the journal of publication, the methods used, the number of hypothesis tests conducted, and the values of t-test statistics. Our objective was to examine the distribution of t-test statistics across different methods, time periods, and journals. Our results indicate that there is no significant difference between the top 5 ranked journals and other journals. Additionally, we observed a downward trend in the prevalence of p-hacking over time. Furthermore, we found that p-hacking and publication bias are associated with the methods used, with instrumental variables (IV) methods exhibiting the most severe issues.

The remainder of this paper is structured as follows. Section 2 will do the data analysis, and it includes the overview of the data, including the data source, the number of variables, and their definition. Section 3 focuses on the analysis of the distribution of these variables. In the end, discussion on these results are proposed in Section 4.

2 Data

I extracted data from the replication package provided by the Brodeur, Cook, and Heyes (2020), including the data set of Table 1 and Figures 1-3. Table 1, Figure 1, Figure 2, and Figure 3 (b, c, d) come from the same data set, while Figure 3a comes from data by another author, BBB. The data in the AAA package is in Stata format, which I read using the Wickham, Miller, and Smith (2023) package in R Core Team (2023) and then stored into a CSV file using Wickham, Hester, and Bryan (2023) .Since I needed to use two different datasets and they couldn't be merged, I stored these datasets separately into different CSV files.

In the data cleaning process, I first extracted the required 5 variables from the raw data, as shown in Table 1. "Journal" represents the name of the published paper. There are a total of 25 Journals. Based on the journal rank, I defined two additional variables for each Journal: "top5" (if the Journal is among the top five, including "Quarterly Journal of Economics," "Journal of Political Economy," "Econometrica," "American Economic Review," and "Review of Economic Studies") and "top3" (if the Journal is among the top three, including "Quarterly Journal of Economics," "Journal of Political Economy," and "American Economic Review"). These variables will be used to analyze the difference in p-hacking by Journal rank. The "Title" variable is used to identify different articles. The variable "t" represents the test statistic in the paper, using methods such as DID, IV, RCT, and RDD.

2.1 Measurement

In the replication package provided by Brodeur, Cook, and Heyes (2020), data was collected from 25 top economic journals for the years 2015 and 2018, encompassing all articles utilizing the DID, IV, RCT, and RDD methods. Table 2 presents a comprehensive list of journals ranked based on the simple impact factor from RePEc. Articles not utilizing the aforementioned methods were disregarded. During sample selection, a rule-based exclusion procedure was employed. For each method, relevant keywords were searched throughout the entire text of published articles to identify the methods used. Articles employing variant sub-methods were manually removed. Notably, articles utilizing matching (DID) or instrumental variables as part of fuzzy RDD were excluded, with a focus on the two-stage least squares (IV) method. Additionally, articles utilizing structural equation modeling were removed. Ultimately, statistical data from 684 articles were collected.

From the included articles, only estimated values were gathered from results tables. For DID, only the main interaction terms were collected unless non-interaction terms were described as coefficients of interest by the authors. For IV, only coefficients of the instrumental variables provided in the second stage were collected. For RDD, estimates of the preferred bandwidth were collected. Preferred bandwidth was determined by reading the text describing the estimates. Specifications such as controlling for third-degree or higher polynomials of forced variables were excluded. Lastly, for articles utilizing multiple methods, estimates were collected for each method separately. For instance, if a paper simultaneously utilized DID and IV, estimates for both methods were collected.

All test statistics "t" in our sample were associated with two-tailed tests. The majority (91%) reported coefficients and standard errors, while others reported z-statistics (4%) or p-values (5%). Since degrees of freedom were not always reported, coefficients and standard errors were assumed to follow an asymptotic standard normal distribution. When z-statistics or p-values were reported, they were converted to equivalent z-statistics.

We also reexamined articles and test statistics from Brodeur et al. (2016) using the same rule-based exclusion procedure, classifying articles by method and retaining only coefficients

Table 1: Variables Definition

variable	Description
journal	Name of the journal
title	Article title
year	Publish year
t	t statistic
method	Used method
top5	Top 5 journal
top3	Top 3 journal

of interest. This yielded 17,518 test statistics from 266 articles published in the top 5 journals from 2005 to 2011. These additional data were used to explore changes in p-hacking over time beyond the 2015 and 2018 samples.

3 Result

Our results are summarized in Table 2 by using Zhu (2021) .We have gathered 21,740 test statistics from 684 articles. DID and IV methods were the most commonly used, with 241 and 281 articles respectively, while RDD was the least utilized, with only 85 articles. On average, each DID article had 24 tests, each IV article had 18 tests, each RCT article had 52 tests, and each RDD article had 37 tests. DID, IV, RCT, and RDD accounted for 27%, 24%, 35%, and 14% of the sample respectively.

Figure 1 a illustrates the distribution of z-statistics across all samples. We used a histogram with a bin width of 0.1. Due to some t-values being excessively large, only data with t-values not exceeding 10 were retained for plotting. Critical values corresponding to significant levels of 10%, 5%, and 1% are marked with green, blue, and red vertical lines respectively. Finally, the density function curve corresponding to the histogram is plotted with a solid black line. The distribution exhibits a bimodal shape: the first peak corresponds to lower z-statistics, while the second peak lies between the green and red lines. There is a noticeable jump to the right of the blue line, indicating a potential presence of p-hacking. Figure 1 b and c depict the distribution of z-statistics divided by journal ranking into top 5 and non-top 5 journals. These three panels are integrated by Auguie (2017) .Both distributions exhibit similar bimodal shapes, suggesting that the degree of p-hacking is unrelated to journal ranking in our sample of the top 25 journals.

Figure 2 displays the distribution of z-statistics for each method. The histogram construction method is the same as in Figure 1. The distribution of DID exhibits two local maxima, with the second local maximum occurring at the critical value corresponding to the 5% significant level. IV shows only one global maximum occurring at the critical value corresponding to the 5% significant level. DID and IV seem to exhibit transitions from statistically insignificant

Table 2: Summary Statistics

journal	DID	IV	RCT	RDD	Articles	Tests
AEJ: Applied Economics	12	13	23	4	46	2242
AEJ: Economic Policy	25	9	5	8	42	1263
AEJ: Macroeconomics	NA	5	NA	NA	5	54
American Economic Review	21	23	14	3	55	1740
Econometrica	2	4	1	4	10	307
Economic Journal	13	22	1	4	38	891
Economic Policy	2	4	NA	NA	6	80
Experimental Economics	NA	2	4	NA	6	79
Journal of Applied Econometrics	NA	4	NA	1	5	86
Journal of Development Economics	13	25	30	3	64	2818
Journal of Economic Growth	2	7	NA	NA	8	100
Journal of Finance	7	15	5	2	27	1135
Journal of Financial Economics	25	16	NA	3	40	635
Journal of Financial Intermediation	7	6	NA	3	16	285
Journal of Human Resources	4	10	5	3	21	752
Journal of International Economics	7	13	NA	1	19	510
Journal of Labor Economics	5	4	8	4	20	653
Journal of Political Economy	4	8	5	2	18	761
Journal of Public Economics	28	18	18	15	74	2605
Journal of Urban Economics	10	16	NA	3	26	660
Journal of the European Economic Association	8	7	6	2	20	491
Quarterly Journal of Economics	5	9	8	6	23	840
Review of Economic Studies	2	3	2	NA	7	306
Review of Economics and Statistics	14	22	10	7	49	1484
Review of Financial Studies	25	16	NA	7	39	963
Total articles	241	281	145	85	684	NA
Total tests	5853	5170	7569	3148	NA	21740

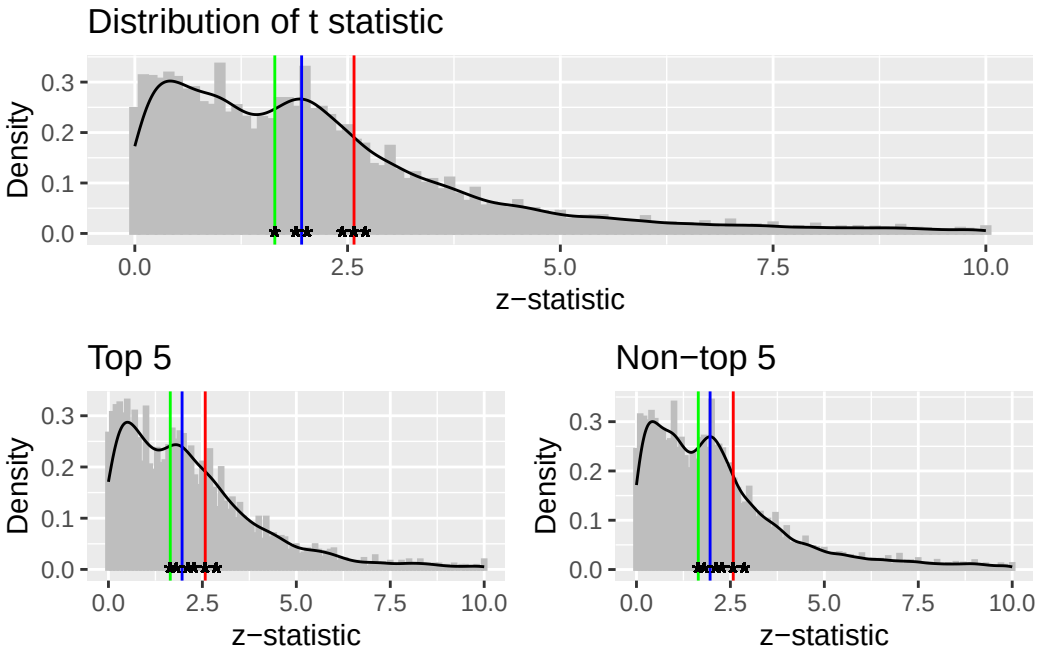

Figure 1: z-statistics in 25 Top Economics Journals

regions to significant intervals. The mismatch in IV appears to be the highest, with a fairly large peak. In stark contrast, RDD presents an almost monotonic decreasing curve, with maximum density close to 0. The distribution of RCT is similar, but with much smaller local maxima close to the critical value at 5% significant level. This suggests that the degree of improper testing allocation in RCT and RDD articles is much greater than in articles using IV and DID methods.

Figure 3 reflects the expected changes in p-hacking over time. Figure 2 (top left) compares tests from three top journals between 2005-2011 and 2015-2018 (top right), while the bottom panel provides a comparison of tests from the top 25 journals in 2015 and 2018. From the two upper panels, it can be observed that for top 3 articles, the distribution between 2005-2011 and 2015 & 2018 is similar, indicating no significant changes in p-hacking for top 3 articles over time. Similarly, comparing the distribution curves for the top 25 in 2015 and 2018, we did not find significant changes over time.

4 Discussion

4.1 harmness of P-hacking and overall discussion

P-hacking, the selective analysis or manipulation of statistical tests to attain desired significance levels, poses a significant challenge in empirical research, especially in fields like

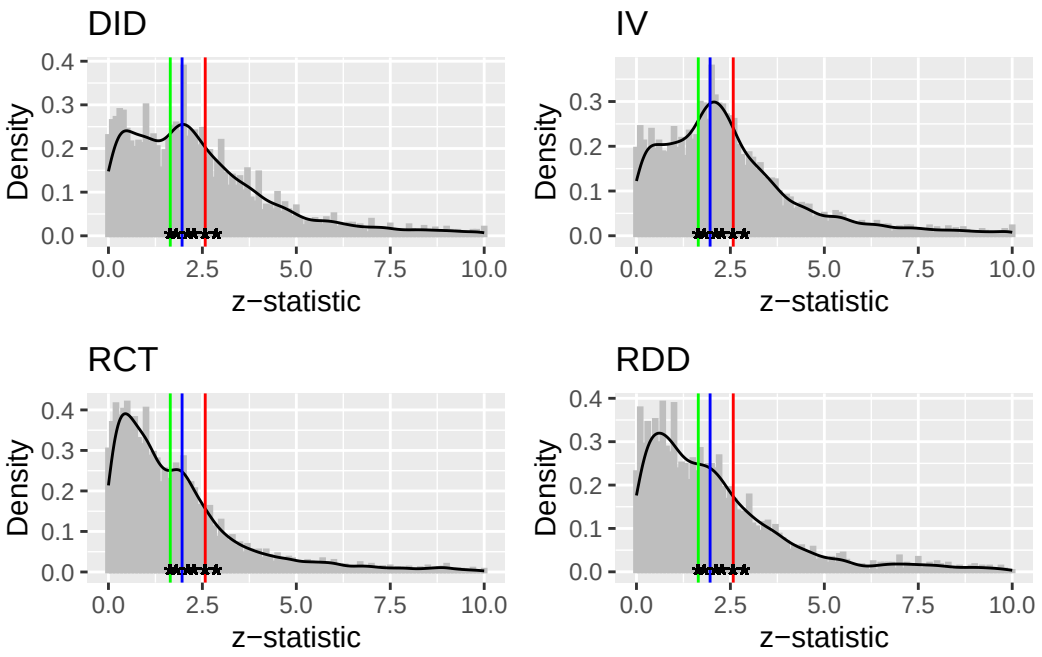

Figure 2: z-statistics by Method

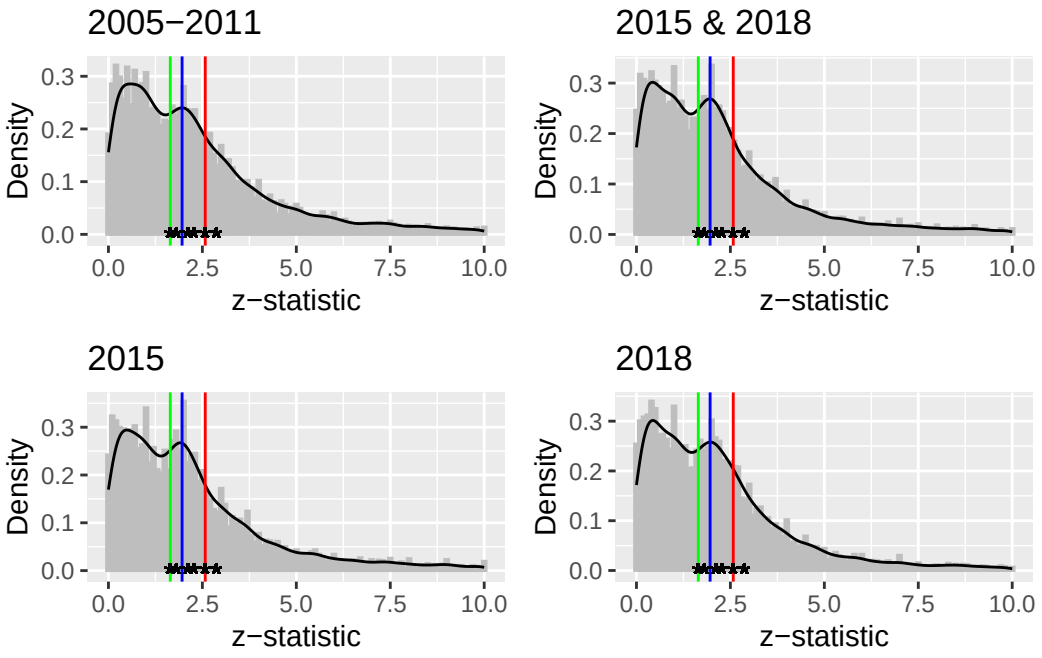

Figure 3: z-statistics over Time

economics where causal inference is pivotal. Its prevalence undermines scientific credibility, leading to false positives and distorting interpretations. Our study delves into p-hacking's prevalence and implications in economic research, particularly in quasi-experimental causal analysis.

Our data collection process was meticulous, ensuring dataset reliability. We extracted data from top economic journals between 2015 and 2018, focusing on articles utilizing quasi-experimental methods like DID, IV, RCT, and RDD. Through a rule-based exclusion procedure and manual review, we aimed to minimize biases and errors in data selection.

Our findings offer insights into p-hacking's scope and trends across methods, timeframes, and journal rankings. We noted a decline in p-hacking over time, hinting at better research practices or heightened methodological awareness among researchers. Notably, we observed variations in p-hacking severity among methods, with IV methods exhibiting more pronounced issues.

Overall, our study sheds light on the complex dynamics of p-hacking in economic research, highlighting the need for continued vigilance and transparency in research practices. By addressing p-hacking and promoting robust methodologies, we can bolster the credibility of economic research and ensure more reliable and accurate findings for informed decision-making.

We have studied how P-Hacking Varies by method, year and journal rank, and I will discuss these topics in the following.

4.2 P-Hacking Varies by Method

The distribution of z-statistics across various causal inference methods unveils nuanced insights into the potential susceptibility to p-hacking in empirical research. While Difference-in-Differences (DID) and Instrumental Variables (IV) methods exhibit transitions from statistically insignificant regions to significant intervals, indicating possible selective reporting for desired significance levels, IV methods notably present a higher mismatch, suggesting potential data manipulation or selective reporting. In contrast, Regression Discontinuity Design (RDD) displays a monotonic decreasing curve, implying a lower likelihood of p-hacking, while Randomized Controlled Trials (RCT) exhibit smaller peaks near critical values, suggesting less severe improper testing allocation compared to IV and DID methods. These findings underscore the importance of robust methodologies and transparent reporting to mitigate the risk of p-hacking and ensure the credibility and reliability of empirical research in economics.

4.3 P-Hacking Varies by Year

The analysis of p-hacking trends over time, as depicted in Figure 3, provides valuable insights into the evolving landscape of empirical research practices. When comparing tests conducted in three top journals between 2005-2011 and 2015-2018, as well as those from the top 25

journals in 2015 and 2018, our findings suggest a relative stability in p-hacking patterns over the years. Specifically, for articles published in the top three journals, the distribution of tests between the two time periods remains similar, indicating no significant changes in p-hacking tendencies for this subset of articles over time. Similarly, the comparison of distribution curves for the top 25 journals in 2015 and 2018 reveals no substantial alterations, further supporting the notion of consistent p-hacking behavior across these years.

4.4 P-Hacking Varies by Journal Rank

The examination of p-hacking tendencies across journal rankings provides valuable insights into the influence of publication prestige on research integrity. In Figure 1, the distribution of z-statistics divided by journal ranking into top 5 and non-top 5 journals reveals similar bimodal shapes for both categories. This finding suggests that the degree of p-hacking remains consistent regardless of journal ranking within our sample of the top 25 journals. Despite the perceived prestige associated with top-tier journals, our analysis suggests that researchers may engage in similar levels of selective data analysis or manipulation of statistical tests across publications of varying rankings. These results underscore the pervasive nature of p-hacking in empirical research, highlighting the importance of promoting transparency, robust methodology, and rigorous peer review processes to uphold the integrity and credibility of scientific publications across all journal tiers. Efforts to address p-hacking should extend beyond journal rank and focus on fostering a culture of research integrity and reproducibility within the scientific community.

4.5 Weaknesses and next steps

One weakness of our study is its reliance on published articles, which may not capture unpublished or rejected research, potentially limiting insights into p-hacking practices. Additionally, our analysis predominantly focuses on the quantitative measurement of p-hacking, primarily examining the distribution of test statistics. This approach may overlook qualitative aspects, such as researchers' motivations and decision-making processes during data analysis. Moreover, our study is confined to economic research, raising questions about the generalizability of our findings to other disciplines. Furthermore, our study only examines articles published in the top 25 economic journals for 2015 and 2018, overlooking contributions from conferences or lower-ranked journals, which could provide a broader perspective on p-hacking practices.

In the future, research could incorporate qualitative methods, such as interviews or surveys, to explore researchers' perspectives on p-hacking and the factors influencing their data analysis decisions. Including unpublished or rejected studies could offer a more comprehensive understanding of p-hacking across different stages of the publication process. Additionally, extending the analysis to fields beyond economics would facilitate a broader exploration of

interdisciplinary p-hacking trends and practices. Lastly, there should be a call for journal reviewers to prioritize fairness over the pursuit of significant results, fostering a more impartial review process.

References

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review* 110 (11): 3634–60.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Evan Miller, and Danny Smith. 2023. *Haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. https://CRAN.R-project.org/package=haven.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.