Disentangling p-Hacking and Publication Bias

Nino Buliskeria

Abstract

This study differentiates p-hacking from publication bias by examining biases resulting from selective reporting within studies versus selective publication of entire studies. Analyzing a dataset of 400 meta-studies, which covers nearly 200,000 estimates from approximately 19,000 individual studies in economics and related social sciences, I observe a notably higher incidence of p-hacking compared to selective publication. Using various meta-regression methods, I find that selective reporting within studies is about 20% more prevalent than publication bias arising from selection among studies. This finding underscores the considerable influence of practices such as p-hacking and method-searching, suggesting that they contribute significantly to selection bias in the economic literature and could affect the perceived reliability of published findings.

JEL Codes: A11, C13, C40
Keywords: selective reporting, publication bias, p-hacking
Acknowledgment: This work was supported by the Charles University Research Center program No. UNCE/HUM/035. This manuscript is part of a project that has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 870245. I thank Jaromir Baxa, Pedro Bom, Ali Elminejad, Alessandro Ferrari, Tomas Havranek, Bob Reed, Tom Stanley, Oliko Vardishvili and Ia Vardishvili for their helpful comments and suggestions. I also thank the participants of the MAER-Net Colloquium at the Universitat de Les Illes Balears and the research seminar at the University of California, Irvine, for their helpful suggestions and comments. The responsibility for all remaining errors and omissions rests solely with me.

1 Introduction

Selective reporting of empirical results can distort our understanding of how robust documented regularities are and give a false impression of their generalizability. Since the early 1980s, the critical examination of empirical research, initiated by Edward Leamer, has catalyzed what is now known as the credibility revolution in economics. This movement has strongly emphasized the importance of meta-research and the replicability of published work.[1] The credibility of empirical research is the cornerstone of scientific progress, yet it remains vulnerable to the influences of p-hacking and publication biases.

Publication bias arises when editorial teams and reviewers prefer studies that demonstrate statistically significant results. Meanwhile, the perception that publication bias is prevalent can lead researchers to abandon studies with unexpected or unpromising results, exacerbating publication bias. On the other hand, p-hacking involves various tactics researchers use, sometimes unintentionally, to achieve more favorable p-values, including "specification search," "p-hacking," or "data dredging" (Brodeur et al., 2020, 2023; Lang, 2023; Mathur, 2022). These tactics can include collecting data until the results appear significant, adjusting econometric models, or setting specific sample criteria to reach desired levels of statistical significance. The urge to engage in p-hacking can come from the perceived importance of statistical significance for the probability of publication (Andrews and Kasy, 2019).

Meta-regression analyzes are widely used to assess the extent of selection bias and to estimate the true population mean, often referred to as "mean-beyond bias" in the literature.[2] These methods generally conceptualize publication bias as a filtering mechanism that impacts a collection of point estimates, which are presumed to be unbiased estimators of the true population effects.[3] However, this foundational assumption is notably vulnerable to selection bias caused by p-hacking, as noted by Irsova, Bom, et al. (2023). The practice of p-hacking, which involves actively seeking specifications that yield significant results, significantly undermines this crucial assumption. p-Hacking can potentially modify both the effect size and the standard error, resulting in spurious pre-

[1] This wave of change has influenced research beyond economics to address what is commonly referred to as the "replication crisis" (Camerer et al., 2018), affecting fields such as medicine and epidemiology with Ioannidis at the forefront (Begley & Ioannidis, 2015; Ioannidis, 2005; Ioannidis et al., 2017), as well as psychology and social sciences. An expanding body of work explores the issues of potential publication biases within economics and various other fields (Andrews & Kasy, 2019; Ashenfelter et al., 1999; Bruns et al., 2019; De Long & Lang, 1992; Doucouliagos & Stanley, 2013; Ferraro & Shukla, 2020; Furukawa, 2019; Havránek, 2015; Ioannidis, 2005; Ioannidis et al., 2017; Leamer, 1983; Miguel et al., 2014; Stanley, 2005, 2008).

[2] There are two primary categories of statistical techniques for detecting and adjusting for publication bias. The first encompasses traditional methods, such as funnel plot analysis and the "incidental" truncation theorem outlined in Greene (1990), which are based on the assumption that results that are statistically significant and align with the desired hypotheses are more likely to be published (Bom & Rachinger, 2019; Duval & Tweedie, 2000; Egger et al., 1997; Furukawa, 2019; Ioannidis et al., 2017; Stanley, 2008; Stanley & Doucouliagos, 2012, 2014). The second category involves modeling the relationship between a study's likelihood of being published and its p-value, thereby defining a parametric structure for the distribution of population effects before selection. Models in this category, such as two-parameter selection models, often show a bias toward the publication of positive results (Andrews & Kasy, 2019; Hedges, 1984, 1992; Iyengar & Greenhouse, 1988; Van Assen et al., 2015; van Aert & Van Assen, 2021; Vevea & Hedges, 1995).

[3] Publication bias is traditionally viewed as a sieve influencing the research submission and publication process, involving decisions made by researchers, journal editors, and peer reviewers. This bias, resulting from study-level selection, is termed "selection across studies" (SAS) by Mathur (2022).

cision (Irsova, Doucouliagos, et al., 2023). Although theoretically the difference between publication bias and p-hacking is distinct, they are observationally equivalent. This observational equivalence challenges the classical metaregression analysis, since it cannot differentiate between the two. The key presumption underpinning the metaregression analysis is the statistical unbiasedness of point estimates and standard errors. The literature acknowledges the consequences of published p-hacked coefficients, but the extent and measurement of p-hacking remain ambiguous. While Brodeur et al. (2023) argue for the dominant role of p-hacking in publication bias, Lang (2023) finds limited evidence for this phenomenon.

The selective publication of significant and large results causes a truncation in the distribution of observed coefficient estimates. As shown in Greene (1990) and elaborated in more detail in Section 2, this truncation leads to a correlation between the observed coefficients ($coef_i$) and their standard errors (SE_i). Through meta-regression analysis, the strength of this correlation (β) is estimated, serving as an indicator of the extent of selection bias [4]:

$$coef_{ij} = \alpha + \beta \cdot SE_{ij} + [\epsilon_i + u_{ij}]$$

Meanwhile, the estimated intercept (α) from this analysis measures the *true mean beyond bias*, adjusted to account for selection bias.

I define p-hacking as the biased selection of the reported point estimate and the standard error pairs within the study, usually by the authors. By controlling for study-specific characteristics, I isolate the bias arising from p-hacking:

$$\text{FE:} \quad coef_{ij} - \overline{coef}_j = \beta^{FE}(SE_{ij} - \overline{SE}_j) + u_{ij}$$

Employing fixed-effects analysis enables the comparison of estimates while canceling the impact of study heterogeneity. By doing so, it becomes possible to identify variations in selection bias that are specifically attributable to variations in within-study coefficient selection, known as p-hacking.

Next, to identify the selection bias between studies, I apply the between-effect estimation on means of coefficient and standard error pairs for each study.

$$\text{BE:} \quad \overline{coef}_j = \alpha + \beta^{BE}\overline{SE}_j + u_j$$

This approach measures the magnitude of selection across studies, the selection type that does not introduce bias in point estimates.

The focus is on five key bias correction estimators: the Egger equation, quantile regression, the Precision-Effect Estimate with Standard Errors (PEESE), the combined PET-PEESE approach, and the endogenous kink model (EK). My objective is to evaluate the extent of selection bias arising from within-study manipulations versus across-study biases. To control for the impressions in meta-regressions coming from the potential presence of the p-hacking, I adopt the instrumental variable approach detailed by Irsova, Bom, et al. (2023) for each estimation technique.

This study also stands out due to its extensive and unique data, encompassing 400 meta-studies that include nearly 200,000 estimates derived from about 19,000 distinct studies. The data for these 400 meta-studies was obtained from the authors when not available in online journal directories (see the Appendix for the list of meta-studies). Next,

[4] Equations in this section are presented for clarity. Please refer to the section 4.2 and 4.4 for further details on theory and application

I combined 412 distinct data sets, synchronizing meta-study and study-level journal titles, and identified the status (working or published article) of the study at the time of meta-study publication (in the journal of online series). Finally, I merged it with a dataset of the SCImago Science Journal Rank on the journal research areas classification to identify the field of meta-study. I base my analysis on this unique and comprehensive data set, which provides a robust platform to examine how biases manifest in published research.

In my analysis of 412 meta-studies, I implement two sets of five key bias correction estimators, each employing an instrumental variable approach. I perform a fixed effect analysis to estimate the extent of bias attributable to p-hacking. Whereas I use a between-effect approach to assess the degree of selection bias arising from selection across studies. This dual approach results in 412 bias estimates for each between- and fixed-effect estimation, which is 4120 regressions in total. To analyze these findings further, I employ a ratio to compare the between- and fixed-effect estimates. Theoretically, as suggested by (Angrist & Pischke, 2009), this ratio, in absolute terms, should be less than one due to the attenuation bias inherent in fixed-effect estimation. However, the median ratio consistently exceeds 1 in all the methodological specifications in my study. My analysis reveals that p-hacking is more prevalent compared to selection between studies, aligned with Brodeur et al. (2023). The results consistently show a higher level of bias in fixed-effect analyzes, indicating a substantial contribution of practices such as p-hacking to selection bias in the economic literature. This outcome indicates a substantial contribution of practices such as p-hacking and method searching to selection bias in the economic literature, leading to a potentially inaccurate perception of robustness in published findings.

The paper is structured as follows. Section 2 discusses the theoretical foundations of bias detection techniques. Section 3 examines the data. Section 4 introduces the empirical techniques and discusses the results. The final section summarizes the findings and implications.

2 Theoretical foundation

According to the traditional definition of publication bias, the research results are selected for publication according to their direction and statistical significance. Although this selective publication process partially truncates the overall distribution of reported results in the literature, in most meta-literature, it is assumed that the chosen results are unbiased estimations of the true underlying effect relative to their respective population. Therefore, most publication bias detection and correction techniques rely on this assumption.

However, Brodeur et al. (2016, 2023), Irsova, Bom, et al. (2023), and Mathur (2022) point to the possible manipulation of design choices that influence standard errors and coefficients to increase the probability of publication. In observational research, the derivation of the standard error is subject to various complicated design choices and with different choices of model specification, both effect size and standard error change. Since both jointly contribute to statistical significance, design choices aiming at increased significance can cause spurious precision and violate the core assumption of unbiased estimates. Violation of this assumption renders meta-regression analysis incapable of correcting for publication bias. Irsova, Bom, et al. (2023) state that in this case *"the simple unweighted mean is often the best, but still no good"*. Although the literature

agrees on the potential consequences of published p-hacked coefficients, the significance of the matter or the way to measure it is ambiguous.

In this section, I discuss the theoretical foundation of metaregression analysis (MRA) and the importance of the underlying assumption of unbiasedness of the point estimate. First, I present the theory behind identifying the true mean beyond bias, then I discuss estimation techniques when the assumption of unbiasdness holds and when it does not. Finally, I show my identification strategy to measure the magnitude of p-hacking compared to selection across studies. For simplicity, I consider a strict rule of selection bias where coefficient estimates that do not satisfy the significance requirement do not get published.[5]

Consider a series of studies that estimate the effect size of a specific research question[6]. Each study uses different sample specifications and techniques to achieve unbiased estimates. In this scenario, the study i estimates an unbiased effect $\hat{\alpha}_i$ expected to be close to the actual true effect, denoted as α_i. The discrepancies between these estimated and true effect sizes result from sampling errors and measurement inaccuracies; therefore $\hat{\alpha}_i$ can be expressed as true effect α_i plus error.

$$\hat{\alpha}_i = \alpha_i + u_i \tag{1}$$

Following the Central Limit Theorem[7], the distribution of the estimated effect size is:

$$\hat{\alpha}_i \sim N(\alpha_i, \sigma_i^2) \tag{2}$$

Furthermore, I follow the conventional assumption that the true effect size follows a normal distribution with a Θ mean and $\aleph^2$ variance[8]:

$$\alpha_i \sim N(\Theta, \aleph^2) \tag{3}$$

This assumption is widely assumed in the meta-research and implies that as the number of studies increases, the distribution of their estimated effects, even with sampling and measurement errors, tends to follow a normal distribution centered around the true effect:

$$\hat{\alpha}_i \sim N(\Theta, \sigma_i^2 + \aleph^2) \tag{4}$$

Therefore:

$$\hat{\alpha}_i = \Theta + u_i \tag{5}$$

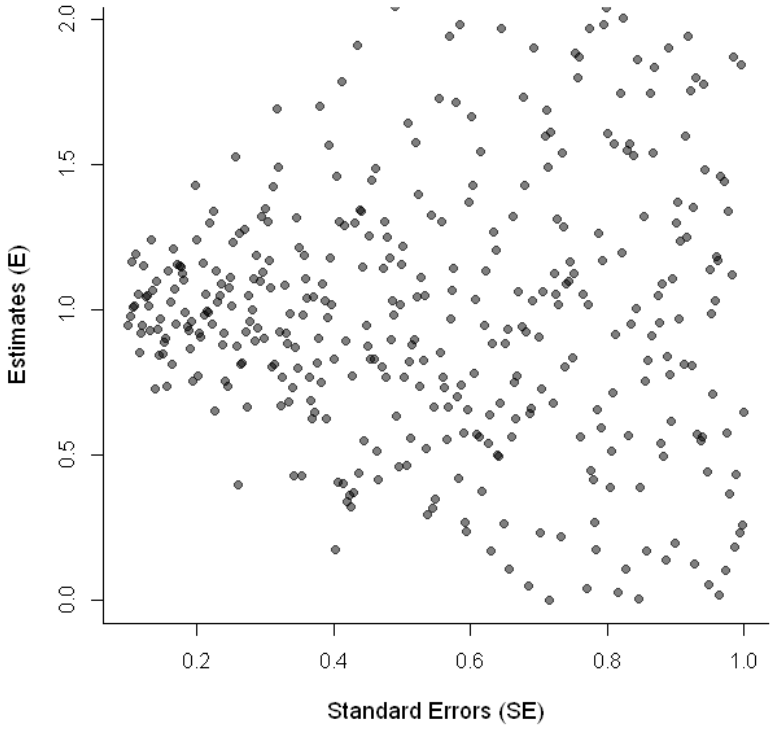
[5]Andrews and Kasy (2019) conclude that studies with a 5% significance level have 30 times higher chances of being published than insignificant results. They estimate the publication probabilities based on replication and meta-analysis approach and provide strong evidence of selectivity based on significance.

[6]Similarly to Jackson and Mackevicius (2023), I start by building the discussion from the point estimates in each study.

[7]The central limit theorem (CLT) states that the average from a random sample for any population (with finite variance), when standardized, has an asymptotic standard normal distribution (Wooldridge, 2002). Here, estimates have not been standardized; therefore, they are normally distributed with mean and variance.

[8]Normality assumption is not essential, here I rather adopt it for ease of demonstration. Most popular meta-analysis techniques assume that the true coefficient estimate, α_i, is statistically independent of its standard error, $sigma_i$, in the population, this easily follows if one assumes that both α_i and $\hat{\alpha}_i$ have the same constant mean Θ across the published studies within a research area. One of the straightforward and most frequently assumed distributions that satisfies the aforementioned requirements in normal distribution

Figure 1: A normally distributed population



where $u_i \sim iid\ N(0, \sigma_u)$ is noise due to the sampling or measurement error, as shown in figure 1.

Let us now consider the classical definition of publication bias. Articles are selected for publication on the basis of their coefficient estimate and significance. This selection criterion leads to missing observations, conditional on coefficient size $\hat{\alpha}_i|\hat{\alpha}_i > a$, and significance level $\hat{\alpha}_i|t_{\hat{\alpha}_i} > c$, where a and c are some constant thresholds. This truncation then creates publication bias (see Figure 2.

The preferences for the coefficient estimate can be in its direction, magnitude, or proximity to conventional beliefs. Let me assume that coefficients larger than some constant a are preferred for simplicity. In the case of truncation based on the coefficient value, only $\hat{\alpha} > a$ are observed; therefore, Equation (4) becomes $\hat{\alpha}_i|\hat{\alpha}_i > a = \hat{\alpha}_i + u|\alpha_i > a$, where $E[u|\alpha_i > a] \neq 0$, and based on (3), to deduct the population mean of true effect Θ bias introduced by truncation needs to be studied:

$$\begin{aligned} E[\hat{\alpha}_i|\hat{\alpha}_i > a] &= \Theta + E[u_i|\hat{\alpha}_i > a] \\ &= \Theta + E[u_i|u_i > a - \Theta] \end{aligned} \tag{6}$$

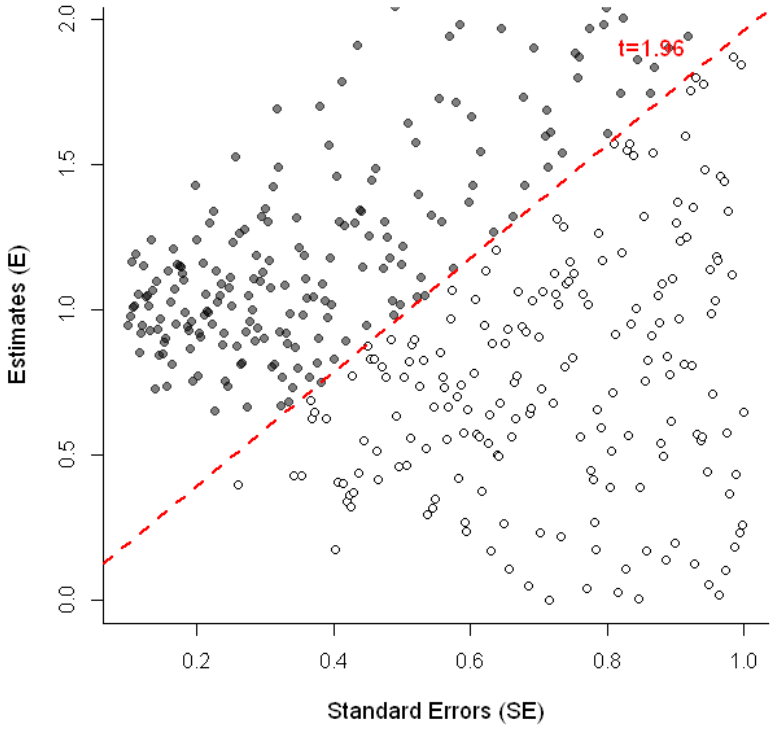
where σ_i is estimated standard error from study i, $E[u_i|u_i > a - \Theta] = \sigma_i\phi(\kappa)/[1 - \Phi(\kappa)]$ and $\kappa = (a - \hat{\alpha}_i)/\sigma_i$ (see Greene, 1990, Theorem 2.2; Wooldridge, 2002; Johnson et al., 1995). Therefore, the conditional expectation of the error term u_i is the product of the estimated standard error and the inverse Mill ratio, which is the ratio of the probability density function to the complementary cumulative distribution function.

$$E[\hat{\alpha}_i|\hat{\alpha}_i > a] = \Theta + \sigma_i \frac{\phi(\kappa)}{[1 - \Phi(\kappa)]}$$

Therefore, the meta-regression is as follows:

$$E[\hat{\alpha}_i|\hat{\alpha}_i > a] = \Theta + \sigma_i\lambda(\kappa) \tag{7}$$

Figure 2: Distribution truncated based on significance, no evidence of p-hacking



Thus, $\lambda(\kappa)$ represents the inverse Mills ratio. If the truncation of the estimated coefficient is above $\alpha_i|\alpha_i < a$, then $\lambda(\kappa) = -\phi(\kappa)/\Phi(\kappa)$.

The truncation of the significance is similar to the truncation of the coefficient estimate, also referred to as incidental truncation [9]. Now, I look at $E[\hat{\alpha}_i|\hat{\alpha}_i/\sigma_i > c]$, where c is the critical value at which the coefficient estimate becomes significant (frequently taken at $c = 1.96$ for the significance level of 5%. To apply the same logic here, it is important to look at the distribution of $\hat{\alpha}_i$ and $\hat{\alpha}_i/\sigma_i$. As discussed above, using CLT, $\alpha_i \sim N(\alpha_i, \sigma_i)$, therefore,

$$\hat{\alpha}_i/\sigma_i \sim N(\alpha_i/\sigma_i, 1) \tag{8}$$

with bivariate normal joint distribution. Therefore, following Theorem 2.5 in Greene (1990)[10]

$$E[\hat{\alpha}_i|\hat{t} > c] = \Theta + \sigma_i\rho\frac{\phi(\kappa_{it})}{1 - \Phi(\kappa_{it})} \tag{9}$$

where $\hat{t} = \hat{\alpha}_i/\sigma_i$, $\kappa_{\hat{t}} = (c - \hat{t})/\sigma_{i\hat{t}}$, and $\rho = corr(\alpha_i, \hat{t}) = 1$. However, considering Equation (7), $\rho = 1$ and $\kappa_{\hat{t}} = (c - \hat{\alpha}_i/\sigma_i)$ result in the same form of meta-regression as shown in Equation (7):

$$E[\hat{\alpha}_i|\hat{t} > c] = \Theta + \sigma_i\lambda(\kappa) \tag{10}$$

To estimate Θ, often referred to as mean beyond bias in the meta-literature, one needs to consistently estimate $\lambda(\kappa)$ first. However, in both cases, the conditional mean is a complex non-linear function of the truncation value σ, α, and λ, while the second term of the equation, $\lambda(\kappa)$, is not constant with respect to α and σ_i. To express the complexity

[9] see in Greene (1990), Theorem 2.5; see Heckman (1979)

[10] first moment of incidental truncation is $\alpha + \rho\sigma\lambda(\kappa_t)$, where ρ is correlation coefficient. However, here $corr(\alpha, \alpha/se) = 1$

of this term, I take the derivative of $E[\hat{\alpha}|truncation]$ with respect to σ, I drop i for simplicity, however, it is assumed as before:

$$\begin{aligned}\partial E[\hat{\alpha}|truncation]/\partial\sigma &= \lambda(\kappa) + \sigma\partial\lambda(\kappa)/\partial\sigma \\ &= \lambda(\kappa) + \sigma\partial\lambda(\kappa)/\partial\kappa \cdot (\partial\kappa/\partial\sigma)\end{aligned}$$

where:

$$\begin{aligned}\partial\lambda(\kappa)/\partial\kappa &= \frac{\phi'(\kappa)[1-\Phi(\kappa)] + \phi(\kappa)\Phi'(\kappa)}{[1-\Phi(\kappa)]^2} \\ &= \frac{\phi'(\kappa)[1-\Phi(\kappa)] + \phi(\kappa)^2}{[1-\Phi(\kappa)]^2} \\ &= -\frac{\phi(\kappa)\cdot\kappa}{[1-\Phi(\kappa)]} + \frac{\phi(\kappa)^2}{[1-\Phi(\kappa)]^2} \\ &= \lambda^2(\kappa) - \kappa\cdot\lambda(\kappa)\end{aligned} \tag{11}$$

as also shown in Heckman (1979). Therefore, after plugging in this derivative and derivative of κ with respect to σ, I have:

$$\partial E[\hat{\alpha}|truncation]/\partial\sigma = \lambda(\kappa) + \frac{\alpha}{\sigma}[\lambda^2(\kappa) - \kappa\cdot\lambda(\kappa)]$$

Equations (7) and (10) is the statistical foundation of the meta-regression model for bias detection, and Equation (2) shows the relation between the expected mean of the truncated estimates and their standard error.

A common approach in the literature to detect bias is to employ a truncated regression model (see Equation 7), also known as the Egger's equation.[11]

$$\hat{\alpha}_i = \alpha + \lambda\sigma_i + \epsilon_i \tag{12}$$

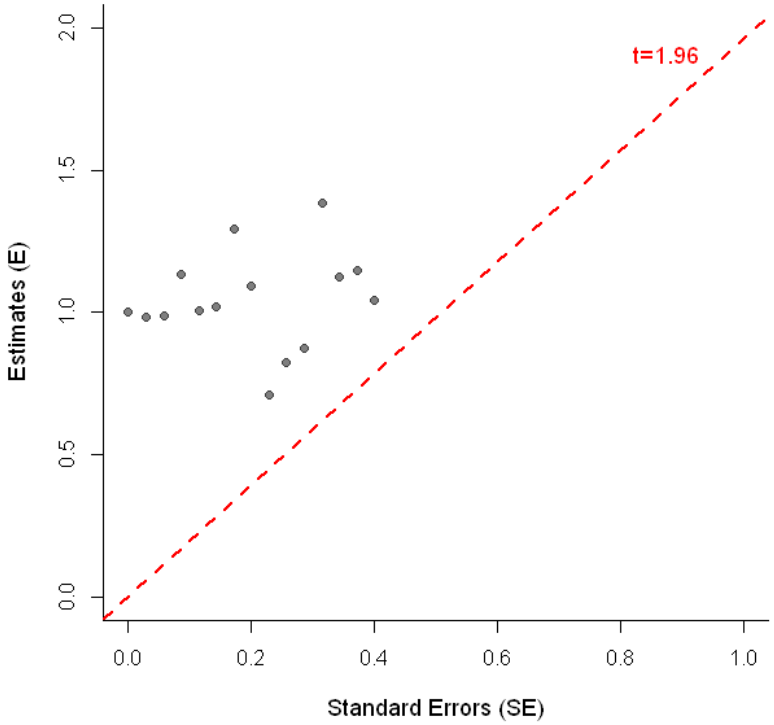
This model aims to determine the presence of bias and to deduce the mean of the target coefficient adjusted for bias from the observed truncated distribution. To alleviate heteroskedasticity, this equation is estimated using weighted least squares, weighted by precision, where t_i is the reported t statistics.

$$t_i = \lambda + \alpha(1/\sigma_i) + u_i \tag{13}$$

The test $H_0 : \alpha = 0$ is known as the *Precision Effect Test* (PET) in the literature and provides a valid test to determine whether there is a nonzero empirical effect after correcting for publication bias (Stanley, 2008). However, Egger's equation struggles to correctly identify the true mean α in cases of nonzero effect size. This is intuitive after comparing Equation (12) with (7), since Egger's regression estimates λ as a constant, while it is a complex function $\lambda(\kappa_i)$ of $\hat{\alpha}$, σ, and the truncation value c, see Equations 11 & 2. Therefore, Egger's equation can correctly measure the extent of bias and identify the mean beyond bias if the underlying empirical effect is zero ($\alpha = 0$), granting the second quadratic term of Equation 2 obsolete - $\partial E[\hat{\alpha}|truncation]/\partial\sigma = \lambda(\kappa)$ and leading to a linear relation between the expected effect and the standard error. However, nonzero cases remain challenging for PET approach.

[11] Frequently written as $coef_i = \alpha + \beta SE_i + u_i$ in the literature, where $coef$ is a coefficient estimate, and SE stands for the standard error. However, here I opted to follow the initial notation.

Figure 3: Study A, no evidence of p-hacking, simulation



In this figure, I present the example of Study A, where there is no evidence of p-hacking since the $t = 1.96$ is not a binding constraint and all results naturally fell on the left side of the line. Hypothetically speaking, study with all naturally significant results would suffer from no selection within study.

The literature strand successfully addresses this issue, using different weighting and Taylor approximation techniques to appeal to the second-order structure of the equation 2 (Bom & Rachinger, 2019; Havránek, 2010; Ioannidis et al., 2017; Stanley, Doucouliagos, et al., 2007; Stanley & Doucouliagos, 2012, 2014). Stanley and Doucouliagos (2014) recommends adopting a quadratic approximation approach, using the weighted least squares (WLS) estimate of the mean beyond bias α.

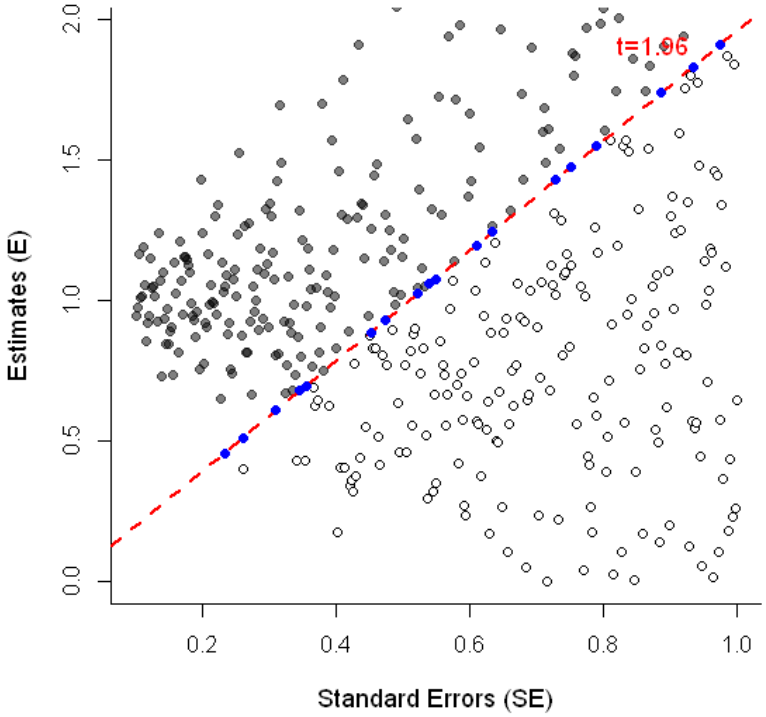
$$\hat{\alpha}_i = \alpha + \lambda\sigma_i^2 + \epsilon_i \qquad \text{or} \tag{14}$$

$$t_i = \lambda\sigma_i + \alpha(1/\sigma_i) + u_i \tag{15}$$

where meta-regression (6) is using $1/\sigma_i$ or $1/\sigma_i^2$ as the weights for the weighted least squared estimation. In the literature, the estimated α is called the *precision effect estimate with standard error* (PEESE) (Havránek, 2010; Stanley, Doucouliagos, et al., 2007; Stanley & Doucouliagos, 2012). Stanley and Doucouliagos (2014) suggest employing the PEESE estimator, Equation 15 only when there is evidence of a nonzero effect (i.e., rejecting $H_0 : \alpha = 0$), and the PET estimator, Equation (12) when accepting $H_0 : \alpha = 0$, which results in the PET-PEESE estimator.

Bom and Rachinger (2019) improve PET-PEESE by proposing the endogenous kink (EK) metaregression model, offering a novel approach to correct for publication bias. A distinctive feature of the EK model is the presence of a 'kink' at a specific cut-off value of the standard error. Below this cutoff point, publication selection is deemed unlikely.

Figure 4: Distribution truncated based on significance, with the evidence of p-hacking



Therefore, the EK model approximates $\lambda(\kappa)$ using a piecewise linear metaregression:

$$\hat{\alpha}_i = \alpha + \delta[\sigma_i - a]I_{\sigma_i \geq a} + \epsilon_i \tag{16}$$

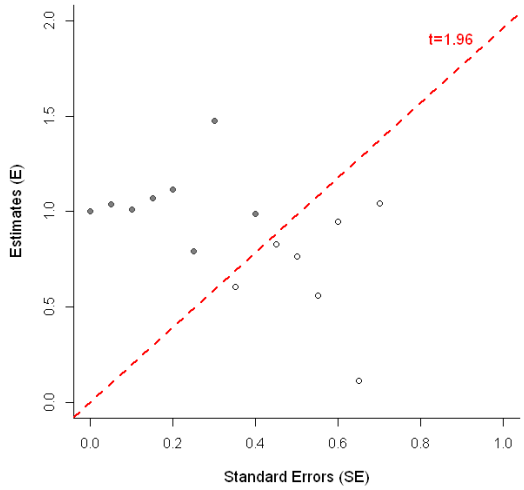
where, $\mathrm{I}_{\sigma_i \geq a}$ is an indicator function that takes the value of one if σ_i is greater than or equal to a, and zero otherwise. Similarly to PET, PET-PEESE, the EK model addresses the heteroskedasticity of $\hat{\alpha}_i$ by dividing each term by $1/\sigma_i$. The EK model endogenously determines the cutoff value based on a preliminary estimate of the true effect and a predefined threshold of statistical significance.

However, the literature is silent on bias detection and correction techniques in the case of spurious precision. All of these methods are based on the implicit belief that the reported nominal precision accurately reflects the true underlying precision. Irsova, Bom, et al. (2023) show that the simple unweighted mean can often outperform complex estimators even when the share of reported spurious precision is very low in the meta-sample. Thus, they argue that when reported standard errors are manipulated conventional solutions, designed to address publication bias, lead further away from true mean. In observational studies, calculating the standard error is often a crucial part of the research process. The process is complex, and varying the computation of confidence intervals will lead the researcher to report different levels of precision for the same estimated effect size, potentially leading to misleading results and spurious precision.

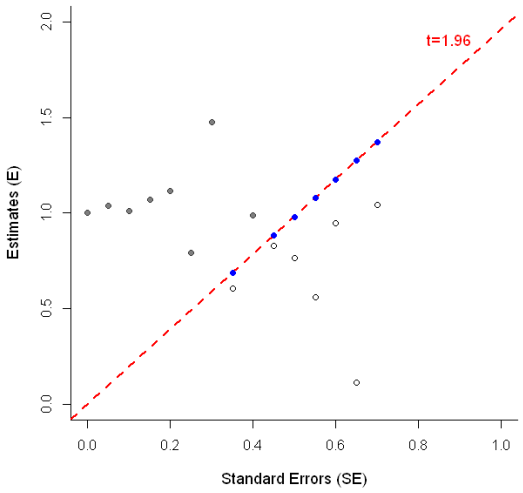
Figure 4 illustrates the distributional consequences of various actions such as cheating, clustering, correcting for heteroskedasticity, and addressing non-stationarity, all undertaken to obtain statistically significant results without a solid theoretical or reasonable basis.

The action of p-hacking can take place in the cases in which researchers increase their selection efforts towards larger estimates in response to noise (larger standard errors) in their data or methods leading to imprecision and insignificance. With these manipula-

Figure 5: Study B, evidence of p-hacking, simulation



(a) Reporting only significant coefficients



(b) Manipulating model specifications to achieve significance

tions, the most precise estimates stay close to the true effect. Therefore, inverse-variance weighting plays a role in reducing bias and improving the efficiency of the aggregated estimate. In contrast, researchers may also achieve statistical significance by reducing the standard error. However, in this case, there is no bias in the reported effect sizes; both the filled and hollow circles would represent identical effect sizes, with the only difference being in precision. The straightforward unweighted average of these estimates is unbiased, but applying inverse-variance weighting would introduce an additional downward bias.

Figure 5 presents the two scenarios of p-hacking, in (a) the author, after conducting a number of estimations and robustness checks, reports only significant results; while (b) shows the case where the author adjusts the specifications of the exercise to achieve significance at the 5% level. The presence of p-hacking introduces the spurious relation between coefficient estimate and standard error, undermining the effectiveness of techniques for detecting and correcting bias.

To control for the spurious relation between estimated coefficients and their standard errors, I use the Meta-analysis Instrumental Variable Estimator (MAIVE) model, where I instrument standard error with the inverse of the sample size[12] , i.e., replace the reported standard error with the portion of the error that can be explained by the sample size. Since in most contexts, the sample size is more difficult to increase than the standard error, the adjusted measure potentially captures the underlying precision better.

$$\sigma_i^2 = \phi_0 + \phi_1(1/n_i) + \nu_i \tag{17}$$

$$\sigma_i = \sqrt{\phi_0 + \phi_1(1/n_i) + \nu_i} \tag{18}$$

where Equation 17 is the first stage regression for the PEESE and Equation 18 for the PET estimation techniques; σ_i is the standard error of the effect size as reported in a primary study; ψ_o is the constant term, n_i denotes the sample size of the primary study, and ν_i is an error term. The error term of the first stage regression, ν_i, absorbs the

[12]here I follow Irsova, Bom, et al. (2023), who offer the MAIVE technique to control for the spurious relation

spurious components of the reported standard error that are attributable to p-hacking. Irsova, Bom, et al. (2023) simulate a realistic p-hacking scenario, suggesting that the MAIVE version of PET-PEESE, without additional inverse variance weights, is more resistant to spurious precision than other existing methods.

The primary objective of the paper is to assess the degree of selection bias resulting from selection within studies (p-hacking) compared to selection across studies (publication bias, file drawer effect). To this end, I plan to conduct my analysis using the instrumental approach as outlined by Irsova, Bom, et al. (2023). My focus is on the five bias correction estimators mentioned above: linear meta-regression, quantile regression, precision effect estimate with standard errors (PEESE), PET-PEESE, and the Endogenous Kink (EK) model. I begin with the linear Egger equation. This is in line with the consensus in the literature that Egger's method is a reliable tool for detecting the presence of selection bias.

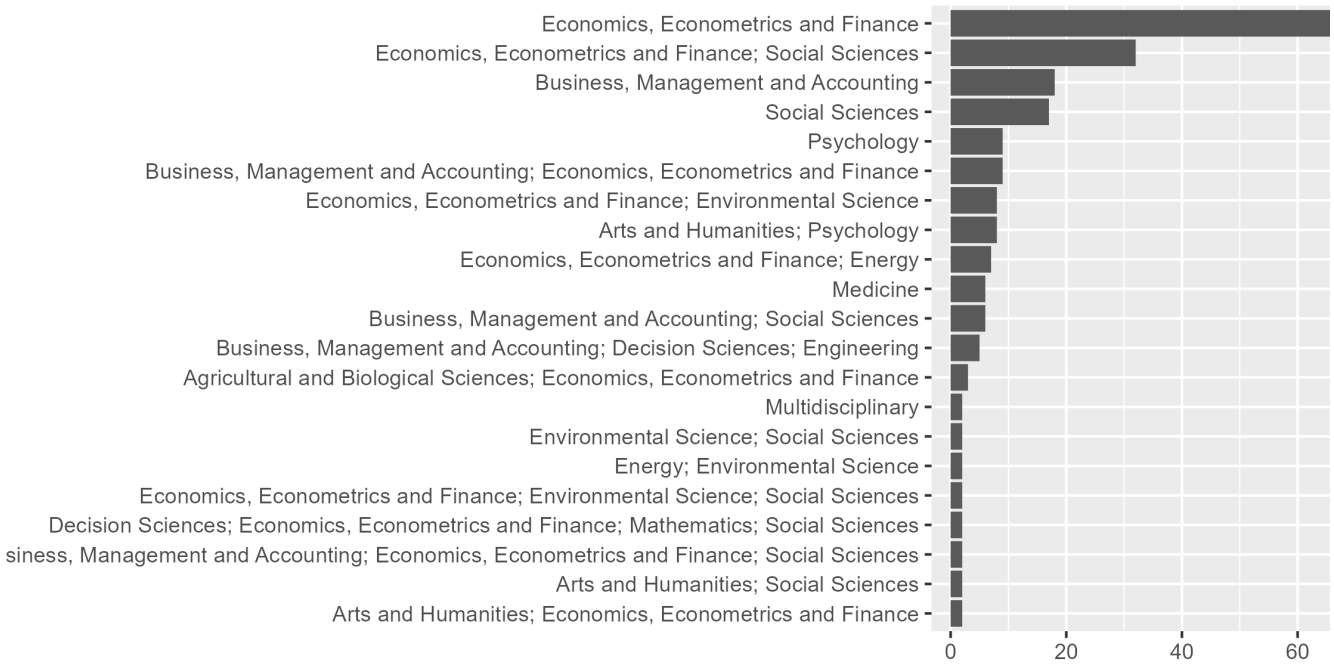
3 Data description

This thesis investigates the sources of selective reporting by examining within-study selection and across-study selection in 400 meta-analyzes, encompassing more than 20,000 studies and 200,000 coefficient estimates from various fields of social sciences, mainly economics. The meta-data set is a collection of data from previous and newly published meta-studies. It contains meta-study and study-level information on authors, titles, publication years, and journals. In addition, the metadata contain coefficient estimates, their respective standard errors, and the sample size of each estimation technique from each study.

Many meta-studies examine closely related questions, often analyzing multiple coefficients of interest corresponding to different true means. In such cases, data from these meta-studies are classified into separate categories and included in the analysis as distinct entities at the meta-level. For example, Balima et al. (2020) analyze the impact of publication selection bias on the macroeconomic effects of inflation targeting. They consider a variety of macroeconomic indicators, including the effects of inflation targeting on inflation, GDP, interest rate volatility, inflation volatility, growth volatility, exchange rate volatility, and deficit. I retain the categorization of Balima et al. (2020)'s data, assigning a unique meta-ID to each category and treating them as independent meta-studies.

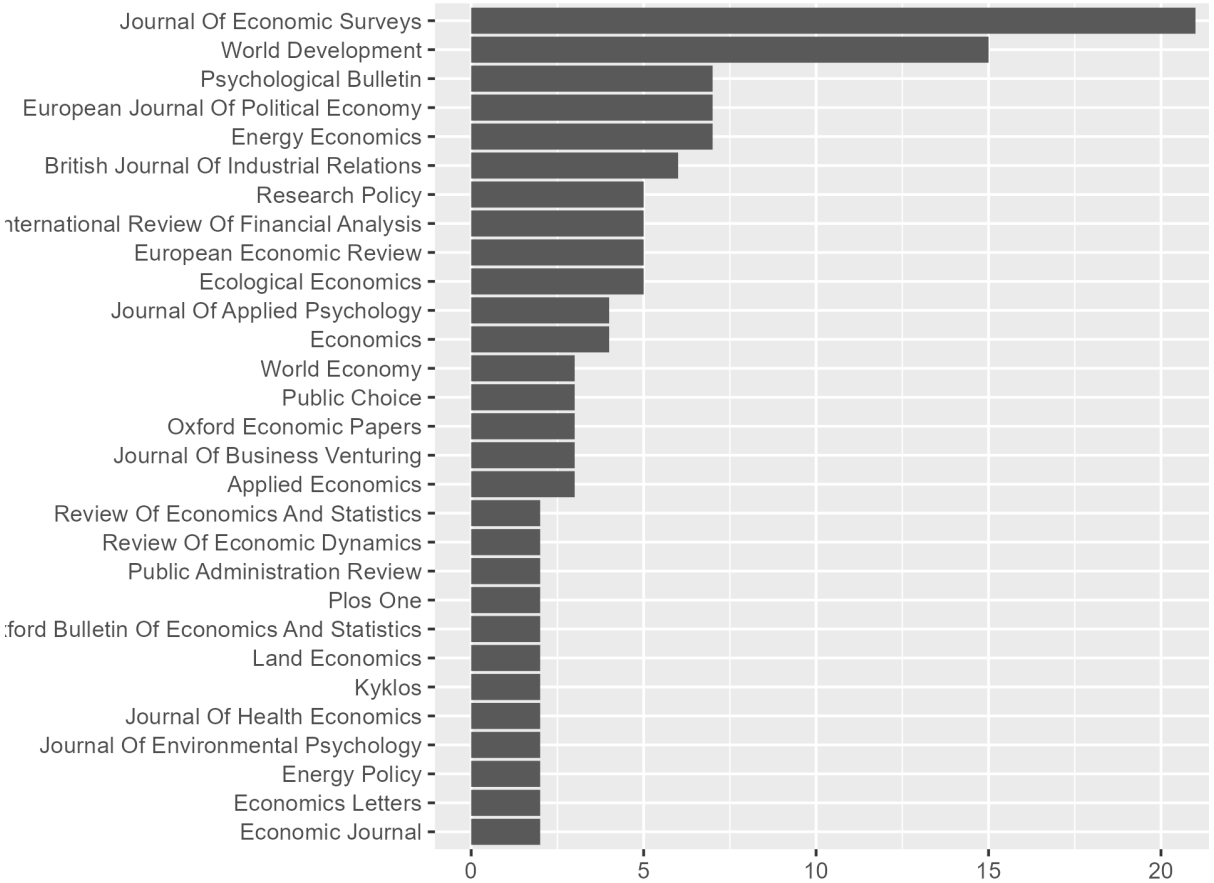
An analysis of the journals where these meta-studies have been published reveals a concentration in various economic disciplines. Figure 6 presents this distribution, categorizing research areas according to the SCImago Journal Rank (SJR). It also shows the frequency of publications within each research area. In particular, the fields of *Economics, Econometrics, and Finance*, with more than 100 meta-analyses, are also mentioned as part of the majority of other area classifications. The repeated appearance of the *Economics, Econometrics, and Finance* classification throughout Figure 6 indicates that our data set mainly comprises estimates drawn from economic research.

Figure 6: The meta-analyses published in journals areas



Note: Journal research areas classification according to the SCImago Science Journal Rank (SJR), https://www.scimagojr.com/journalrank.php?area=2000

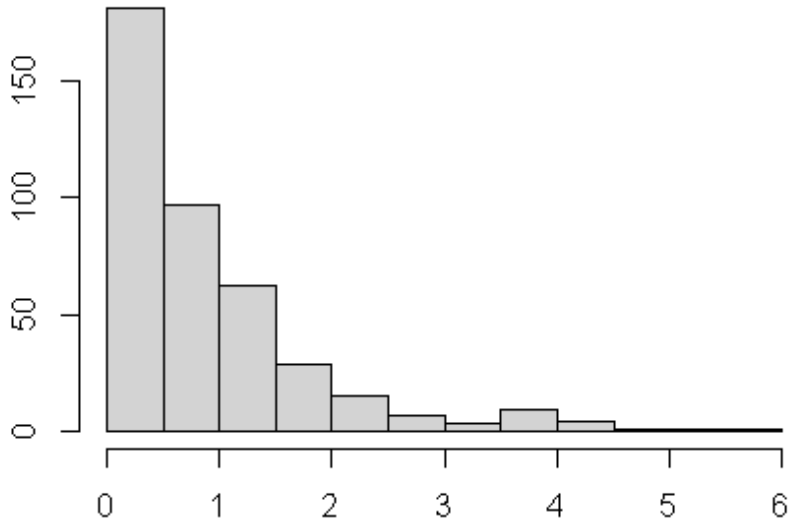
Figure 7: Meta-analyses per journal



Note: a list of journals that are the most frequent publishers of meta-studies included in the dataset.

Figure 6 shows the journals that most frequently publish meta-analyses in the data. Not surprisingly, it reflects the picture that can be seen in Figure 6, where the most frequent research area is economics. In Figure 7, it is apparent that these meta-studies are published more frequently in economic outlets, sometimes psychology, or in interdisciplinary journals such as *Journal of Health Economics*. I present only those journals that have published meta-study in the sample at least twice; however, similarly to Figure 6, the economic journals are the majority of the journals, and social science and interdisciplinary journals are the second most frequent and rarely medicine.

Figure 8: Distribution of Selectivity in Empirical Economics.



Note: Bias estimated from Egger's regression, $coef_i = \alpha + \beta SE_i + \epsilon_i$. The bias is considered *small to modest* if $|\beta < 1|$, *substantial* if $1 \leq |\beta| \leq 2$, and *severe* for $|\beta > 2|$. I find *substantial* selectivity across 91 different topics and *severe* in 44 topics in economics & social sciences. For 278 areas, bias falls in the little to modest category.

To understand the extent of bias in the literature, I use Egger's regression $coef_{ij} = \alpha + \beta SE_{ij} + \epsilon_{ij}$, where $coef_{ij}$ & SE_{ij} is the estimated coefficient and standard error pair j of study i, α is the mean beyond bias, β estimates the extent and existence of bias. I run this regression analysis separately on data from k meta-studies, obtaining the k number of β coefficients for each topic. Figure 8 shows the distribution of β_k on different topics. Doucouliagos and Stanley (2013) categorizes the biases in *little to modest* category if $|\beta < 1|$, *substantial* if $1 \leq |\beta| \leq 2$ and *severe* for $|\beta > 2|$. I find *substantial* selectivity across 91 different topics and *severe* in 44 topics in economics & social sciences. For 278 areas, bias falls into the little to modest category.

Finally, in Figure 9, I look at the distribution of t-statistics in published articles and show evidence of potential p-hacking, as discussed in Brodeur et al. (2023). I use the de-rounding technique and weight the z-statistics (measured as $coef_{ij}/SE_{ij}$) with the inverse of the number of tests present in each article and superimpose an Epanechnikov kernel density curve on the histogram. De-rounding does not change the shape of the distribution; it only smooths potential discontinuities in histograms. Figure 9 presents the two-humped camel-shaped pattern, bunching at $z = 1.96$, indicating the existence of p-hacking. However, as pointed out in Kranz and Pütz (2021), this approach cannot

Figure 9: De-rounded & weighted distribution of z-statistics of published papers.

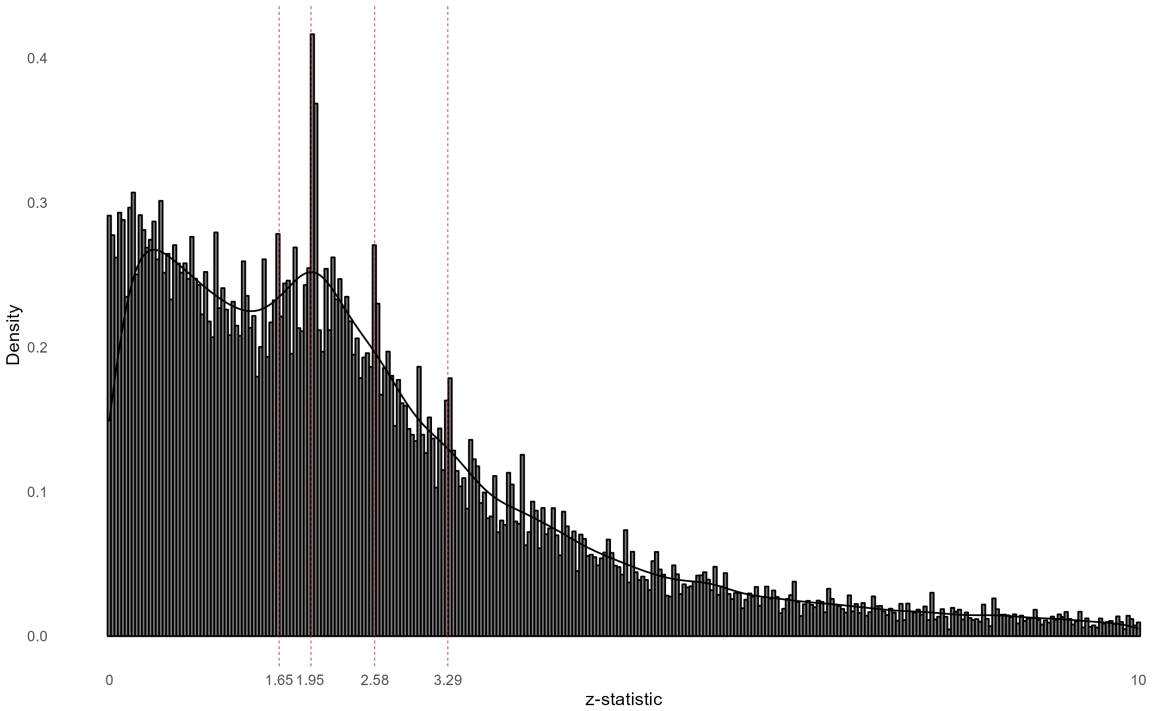


Figure is showing the distribution of z-statistics of coefficient estimates in the published papers. The distribution is de-rounded to control for the Note: The two-humped camel-shaped pattern, similar to Brodeur et al. (2020, 2023), is evident. I superimpose an Epanechnikov kernel density curve.

explain the excess share of observed z-statistics near zero.

The observed distribution of z-statistics, even adjusted for rounding, consistently shows two distinct peaks, one at zero and one around $z = 2$, Figure 9. However, Kranz and Pütz (2021) point out that this second peak does not necessarily indicate p-hacking or publication bias. It could also be explained by a latent mixed distribution resulting from varying research objectives. For example, some studies could refine previous findings with significant effects, while others could be more exploratory, lacking a solid prior assumption of the actual effects being present. To demonstrate this numerically, Kranz and Pütz (2021) consider 5,000 random samples from a combination of three Cauchy distributions, each with a scale parameter of 0.8: one distribution has a center at 0, representing exploratory research, while the other two, centered at -2 and 2, represent more focused research. They show that the resulting distribution of absolute z-statistics is very similar to the empirical distribution in the pooled data in Figure 9. This paper contributes to this discussion by analyzing similar questions based on metaregression analysis.

4 Estimation and results

There should be no correlation between estimates and standard errors if there is no publication bias, that is, selection within (SWS) or across studies (SAS). Therefore, for now I assume that any correlation between the coefficient $coef_{ij}$ and its standard error

SE_{ij} indicates the existence of bias. Therefore, the correlation between $coef_{ij}$ and SE_{ij} within the study indicates bias from SWS, and the correlation between the mean study estimates indicates bias due to SAS[13]. I run 800 regressions to estimate bias coefficients for each research question and separately evaluate the extent of the selection of the results coming from the within-study and between-study variation.

I estimate the extent of selection for each meta-analysis k, study j, and estimate i, using the following meta-regression:

$$coef_{ij} = \alpha + \beta SE_{ij} + e_j + u_{ij} \tag{19}$$

Where $coef_{ij}$ is the coefficient estimate i of the study j; SE_{ij} is the corresponding standard error; e_j indicates characteristics specific to the study and u_{ij} is the error term. This regression cannot differentiate between the selection within- and between-studies, however, it can serve as a benchmark for the comparison. Meta-regression of this type is most frequently used in the literature; however, there can be two issues that present the problem of identifying the estimated β as a measure of selection bias as a whole. First, it is implausible that the pairs of ($coeff_{ij}$; SE_{ij}) and ($coeff_{kj}$ and SE_{kj}) are independent. This assumption can be relaxed if one assumes that the authors and editors select each coefficient estimate independently and separately.[14] However, if the researcher is involved in p-hacking, then the assumption that each coefficient estimate was selected on its own merit is implausible. The second problem arises when one considers the existence of p-hacking, since the necessary assumption that estimated standard errors are unbiased SE_{ij} is also unlikely, therefore, equation 19 suffers from the spurious correlation and cannot accurately estimate the extent of selection bias β in the literature. To address this issue, I use the Meta-analysis Instrumental Variable Estimator (MAIVE) and instrument standard errors using the respective sample size in the first stage to replace the reported standard error, SE_{ij}, with the portion of the error that can be explained by the sample size. Irsova, Doucouliagos, et al. (2023) argue in favor of using the sample size as an instrument for reported standard errors. The reported variance (SE^2) is a linear function of the inverse of the sample size used in the primary study by definition. The sample size is not estimated, so it is free from measurement error. Changes in methodology generally have no effect on the sample size and neither do the choice of control variables. The sample size appears to be more resistant to selection bias, as gathering additional data is more challenging than manipulating the standard error to reach significance. Endogeneity might still persist if researchers, anticipating smaller effects, opt for larger experiments. However, in the context of observational studies, researchers generally use all available data.

To isolate the bias coming from within-study selection, I need to control the study-specific characteristics. I do this by applying fixed effects estimation, demeaning the estimates by the study mean effect and mean standard error:

$$\text{FE:} \quad coef_{ij} - \overline{coef}_j = \beta^{FE}(SE_{ij} - \overline{SE}_j) + u_{ij} \tag{20}$$

The fixed effect estimator takes care of the fixed effect of e_j for the unobserved study by subtracting the mean estimates of the study. This approach allows me to estimate the measure of bias, $\hat{\beta}^{FE}$, coming from the within-study variation.

[13]The caveat here is that coefficients within study are less likely to be independent, however when controlling for the fixed effects, in case of SWS, and taking mean estimates, in case of SAS, this issue should resolve.

[14]see Andrews and Kasy (2019) for more detailed discussion.

Next, to study the extent of publication bias, I look at the extent of selection between studies. Here, I need to proxy a selection criterion for each study - ideally, it would be a main result or a set of results based on which the paper was selected for publication. Unfortunately, I do not have information on which of the estimates is more important in the pool of reported estimates. Therefore, I revert to taking mean estimates as the average story told in the manuscript and the average criteria based on which the publication decision is made.

$$\text{BE:} \quad \overline{coef}_j = \alpha + \beta^{BE}\overline{SE}_j + u_j \tag{21}$$

Therefore, I study the variations between studies using the averages of the estimates for each study.

Finally, with similar rationality, I employ the PEESE, PET-PEESE, and EK model approaches to consistently estimate the extent of selection bias. As above, I run these regressions on demeaned reported estimates first and mean estimates second, to analyze the extent of selection bias that arises from selection within the study and between the studies, respectively.

Figure 10: Different types of selection biases influencing published work

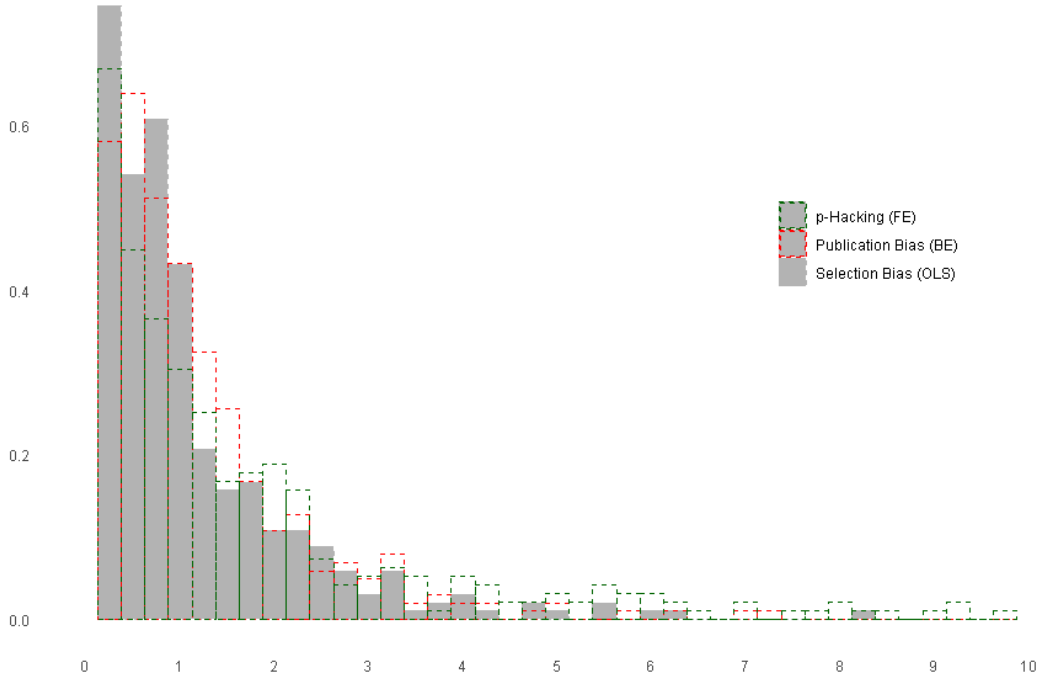


Figure presents the distribution of estimated $\hat{\beta}$ from fixed effect, between effect and OLS estimations, where β^{FE} is extent of within study selection - measure of p-hacking, β^{BE} measures the extent of publication bias defined as selection across study, β^{OLS} estimates the average selectivity in the literature and is the most common version of the meta-regression. Note that these results are retrieved from analysis of Published Paper sub-sample.

4.1 Selection within vs. across study

The Figure 10 shows the distribution of β coefficient from the Fixed-effect (20), between-effect (21), and OLS (19) estimated for 400 subsamples separately. The distribution of

the coefficient β estimated from the OLS regressions, presented as the gray shadow in the figure, is the average effect of selection in the published literature. The measure of bias from the within-study variation indicates the extent of p-hacking (in green); and the measure of bias coming from the between-study variation indicates the extent of publication bias (in red). In Figure 10, when looking at part of the distribution that shows little or no bias $|\beta| < 1$, as well as the moderate level of bias $1 < |\beta| < 2$, the selection between studies seems to be more relevant. But as the severity of the selection bias increases, p-hacking plays a larger role in the selection bias.

Figure 11: Distribution of $\Psi_k = |\beta_k^{FE}/\beta_k^{BE}|$

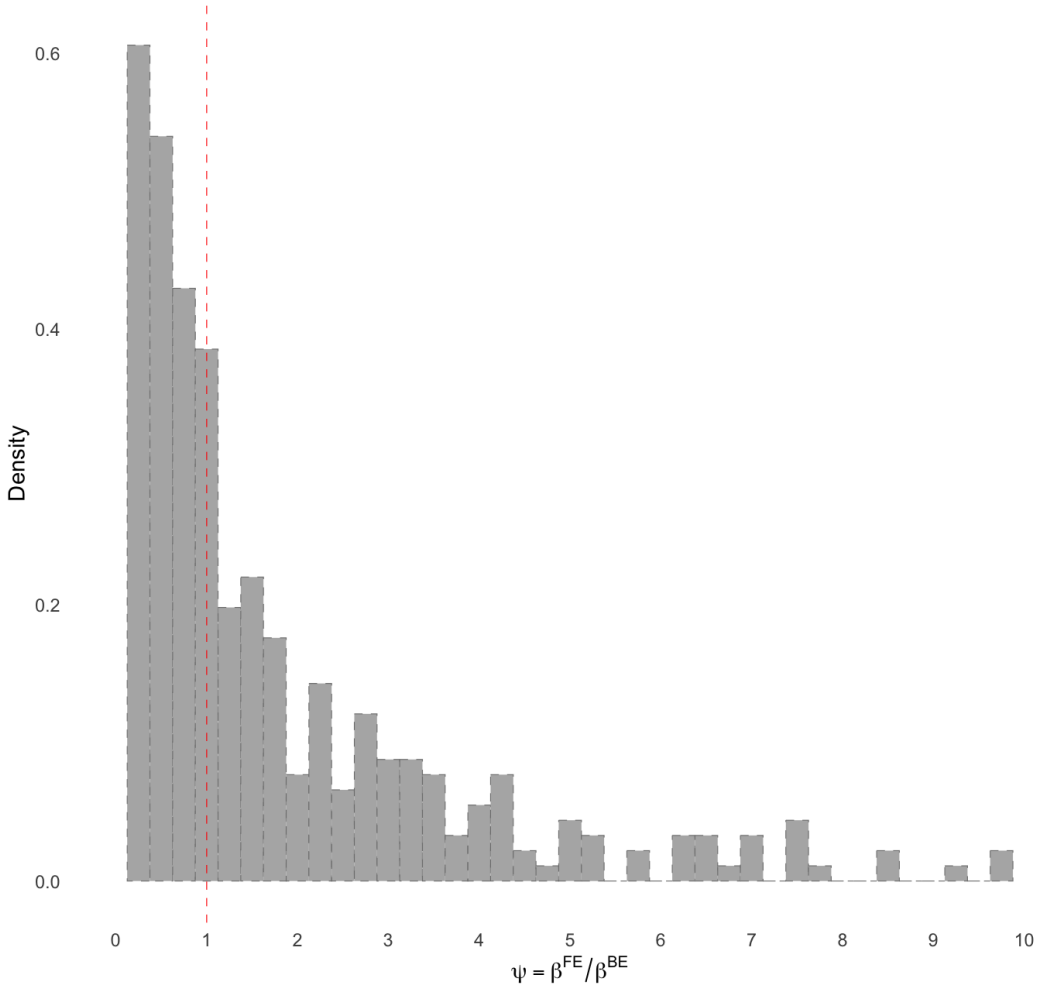


Figure shows comparison of within $|\beta_k^{FE}|$ and between $|\beta_k^{BE}|$ selection using ratio.

Finally, I calculate β_k^{FE} and β_k^{BE} and derive $\psi_k = \beta_k^{FE}/\beta_k^{BE}$ for each meta-study k based on the subsample of published results. Figure 11 shows the distribution of ψ_k with a significant part of the distribution on the right side of red line indicating threshold where $\beta_k^{FE} > \beta_k^{BE}$ has a long tail.

I estimate the ψ_k ratio from the fixed effect and between the effect models[15] and I present the median and mean values of ψ_k with the 95% confidence interval (CI) con-

[15]winsorized on 1, 2.5, and 5%. Table 1 shows the results of the most liberal 1% winsorization. However, 2.4% and 5% winsorization showed very similar results.

structed using t statistics for mean and bootstrapping with a sample with multiple repetitions for the median. Next, to alleviate the effect of outliers, I apply median regression on the original data without winsorization. The both results are consistent in that, they both predict over 10% larger effect of p-hacking compared to the publication bias in the bias caused by selection of the results for publication. Next, in Table 2, I show the analy-

Table 1: Selection within vs. across study, published papers

	Linear Regression	**Quantile Regression**
Median	1.18	1.11
	[1.03; 1.48]	[0.96; 1.28]
Mean	7.78	9.52
	[5.13; 10.44]	[4.31; 14.73]
Number of Meta-Studies	409	407

In the table, the median and mean values of ψ_k are detailed, each accompanied by a 95% confidence interval (CI). These intervals are calculated using the t-statistics for the mean and using bootstrapping with multiple repetitions for the median. Additionally, the data set has been winsorized at the 1st and 99th percentiles to enhance its statistical robustness. The data set comprises estimates exclusively from published papers.

sis based on PEESE, PET-PEESE, and EK regressions. To control for possible p-hacking and more accurately estimate the extent of biased selection, I instrument the reported standard errors, SE_i, in the first stage [16] with the inverse of the sample size to the instrument for the standard errors. In Table 2 I report the median and means of estimates that show strong correlation on the first stage as evidence of instrument's relevance.

Table 2: Selection within vs. across study, published papers

	PEESE	**PET-PEESE**	**EK**
Median	1.33	1.29	1.22
	[1.15; 1.51]	[1.05; 1.76]	[1.07; 1.44]
Mean	7.44	7.58	4.41
	[1.66; 13.22]	[1.91; 13.25]	[2.66; 6.17]
Number of Meta-Studies	191	191	191

In this table, the median and mean values of ψ_k are presented, derived from the Instrumental Variable (IV) regressions of the PEESE, PET-PEESE and EK models. These values are accompanied by 95% confidence intervals (CIs), which are constructed using t-statistics for the mean and bootstrapping with multiple repetitions for the median. The dataset has been winsorized at the 1st and 99th percentiles. The number of meta-studies included in this analysis has been reduced to 206, as ψ_k values from regressions with first-stage F statistics less than 10 have been excluded. The data set comprises estimates exclusively from published papers.

In all five approaches (Tables 1 & 2), I find that the bias arising from the variation

[16] suggestions Irsova, Bom, et al. (2023)

within the study is greater than the selection between studies. Although the mean value is greater than 5 in all cases, this is probably due to the long tails of selection bias and ration ψ_k, see the figures 10 and 11. Therefore, looking at the median value of ψ_k is essential. Together, the median and mean values of the ratio suggest that selection within studies is consistently larger compared to selection across studies, pointing to the prevalent evidence of practices like method searching and p hacking in the published literature.

Table 3: Selection within vs. across study, all papers

	Linear Regression	**Quantile Regression**
Median	1.16	1.12
Median CI	[1.06; 1.46]	[0.97; 1.38]
Mean	7.85	8.84
Mean CI	[4.84; 10.87]	[1.63; 16.06]
Number of Meta-Studies	412	368

In the table, the median and mean values of ψ_k are detailed, each accompanied by a 95% confidence interval (CI). These intervals are calculated using t-statistics for the mean and bootstrapping with multiple repetitions for the median. Additionally, the data set has been winsorized at the 1st and 99th percentiles to enhance its statistical robustness.

These conclusions are drawn from looking at the published results. Next, I look at a complete dataset that contains results from published papers and working papers to evaluate the comparison of selection within and across studies in general.

Table 4: Selection within vs. across study all papers

	PEESE	**PET-PEESE**	**EK**
Median	1.21	1.28	1.28
Median CI	[1.12; 1.44]	[1.10; 1.82]	[1.08; 1.51]
Mean	8.33	7.02	4.45
Mean CI	[2.21; 14.44]	[1.73; 12.31]	[1.93; 6.96]
Number of Meta-Studies	206	206	206

In this table, the median and mean values of ψ_k are presented, derived from the Instrumental Variable (IV) regressions of the PEESE, PET-PEESE, and EK models. These values are accompanied by 95% confidence intervals (CIs), which are constructed using t statistics for the mean and bootstrapping with multiple repetitions for the median. The data set has been winsorized at the 1st and 99th percentiles. The number of meta-studies included in this analysis has been reduced to 206, as the ψ_k values of regressions with first-stage F-statistics less than 10 have been excluded.

However, Tables 4 and 5 demonstrate that the findings derived exclusively from the published literature are consistent with those obtained from the entire data set. The Selection Within Studies (SWS) is consistently found to be more pronounced than Selection Across Studies (SAS). This pattern reinforces the notion that significant selection

occurs at the research stage, indicating a tendency to report certain results while omitting others, potentially to strengthen the researcher's argument or narrative.

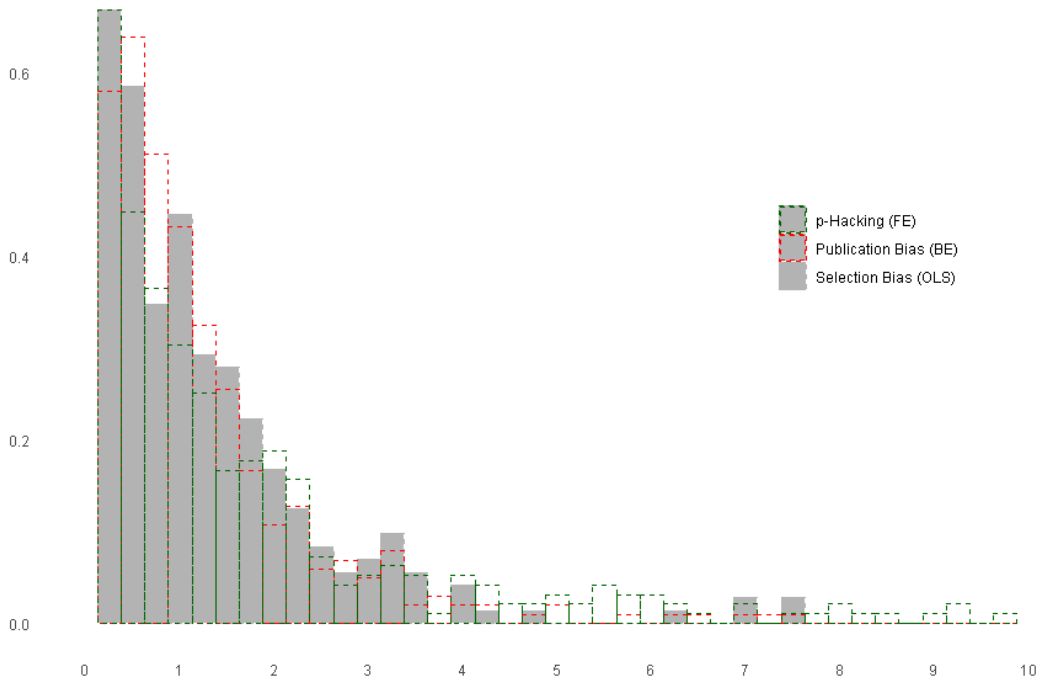
The patterns of selection across and within studies are repeated when analyzing the whole dataset consisting of over 15000 published and 3500 working papers. Next, I look at the selection bias in working papers in comparison to published papers.

4.2 Working papers vs published papers

To understand how to correct and potentially prevent selection bias within and between studies, it is important to explore the stages at which selection occurs. Selection across studies may occur at the submission and revision stage, or much earlier, when the researcher decides whether or not to write the paper. Moreover, while previous results have suggested the existence of a significant level of within-study selection, understanding the effect of the publication process on p-hacking is crucial. To this end, in this section, I first investigate the extent of within and between study selection in a working paper subsample, comparing these two types of biases. Subsequently, I compare the within and between study selection in working papers with that in published papers.

Figure 12 shows the distribution of selection bias, p-hacking and publication bias in working papers. In the realm of working papers, "publication bias" should be viewed as the decision by researchers to write the paper after receiving initial results or not. The phenomenon in which the research chooses to write the research paper according to the obtained results is frequently referred to as a "file-drawer problem" in the literature. Here, also, selection across studies dominates for the low selectivity in reported results,

Figure 12: Selection bias in working paper $|\beta_k^{WP}|$ subset.



and as the selection bias becomes more severe in different fields of research, the effect of selection within study becomes more prominent. To compare the effect of the publication process on bias, I perform a similar analysis as before and compare the extent of these selection biases in the results reported in the working and published articles, see Table 5.

In Table 5. I have reported results from linear, quantile, PEESE, PET-PEESE, and endogenous kick model estimations. As before, the last three use the instrumental variable approach to control for the spurious relation caused by the existence of p-hacking. The first section of the table shows the medians of the $\Psi_k = \left|\beta_{WP;k} \Big/ \beta_{P;k}\right|$ ratio comparing the average selection bias in the results of the working and published papers. Although linear estimations show larger selectivity in the results reported in the working papers, non-linear estimation models do not show such a large difference.

Next, to explore the question of whether the publication process accelerates or reduces selection, I look at the within- and between-study selection comparison separately. Comparison of p-hacking in the working and published papers shows that within-study selection is significantly larger in the results reported in the working papers. In contrast, there are no significant differences in the selection between studies in published papers compared to working papers.

The results in Tables 1, 2 and 5, show that the p hacking dominates compared to the publication bias in published research; however, published results suffer from less within-study selection compared to working papers. Table 5 shows on average greater evidence of p-hacking in working compared to published papers. Therefore, I conclude that the publication process filters out a significant portion of p-hacked results.[17]

These results highlight the widespread nature of selection biases in academic research. The upper section of Table 5 shows that the decision to write a research paper suffers from a selection bias similar to the journal's decision to publish. In essence, the biases affecting what gets written are strongly mirrored in what gets published. However, the primary driver of this phenomenon remains unclear, whether it is shaped more by the anticipations and decisions of journals and editors, or by researchers' beliefs about what is likely to be accepted. On the one hand, researchers could potentially correctly foresee the publication potential of their work and choose not to draft a manuscript that has a lower chance of acceptance. On the other hand, they might only submit manuscripts that they *believe* to likely be published, thereby limiting the array of choices available to journals, creating a self-fulfilling prophecy: even if journals exhibit no selection bias, they end up publishing only a partial narrative because they receive a non-representative sample of research outcomes. However, these results also point to the mitigating role of the publication process in the selection of estimates *within* the study. Table 5, middle section shows that selection within study dominates in working paper sub-sample, leading me to believe that significant portion of p-hacking is filtered before the studies are published.

[17] This conclusion is inline with the findings in Brodeur et al. (2023).

Table 5: Comparison of biased selection in working and published papers

	Linear	**Quantile**	**PEESE**	**PET-PEESE**	**EK**
Selective Reporting $\Psi_k = \left\lvert \beta_{WP;k} / \beta_{P;k} \right\rvert$					
Median	1.23	1.27	1.02	1.13	1.08
	[1.05; 1.55]	[1.06; 1.61]	[0.86; 1.22]	[1.00; 1.44]	[0.88; 1.21]
Meta-Studies	269	284	187	186	152
p-Hacking, Selective Reporting *within* study, $\Psi_k^{FE} = \left\lvert \beta_{WP;k}^{FE} / \beta_{P;k}^{FE} \right\rvert$					
Median	1.16	1.76	1.31	1.67	1.12
	[0.86; 1.28]	[1.36; 2.11]	[0.90; 1.74]	[1.12; 2.32]	[0.99; 1.68]
Meta-Studies	194	282	169	169	169
Publication Bias, Selective Reporting *between* studies, $\Psi_k^{BE} = \left\lvert \beta_{WP;k}^{BE} / \beta_{P;k}^{BE} \right\rvert$					
Median	1.16	1.34	0.93	1.05	0.97
	[0.86; 1.29]	[1.13; 1.66]	[0.74; 1.07]	[0.85; 1.24]	[0.86; 1.07]
Meta-Studies	195	288	134	134	134

This table shows the comparison of biased selection in working papers and published papers. For this, I show the median values of $\Psi_k = \left\lvert \beta_{WP;k} / \beta_{P;k} \right\rvert$; while, Ψ_k^{FE} compares the extent of *p*-hacking and Ψ_k^{BE} compares the extent of publication bias in working and published papers. In the columns (1) & (2), the median and mean values of ψ_k are detailed, each accompanied by a 95% confidence interval (CI). These intervals are calculated using the *t*-statistics for the mean and using bootstrapping with multiple repetitions for the median. Additionally, the data set has been winorized at the 1st and 99th percentiles to enhance its statistical robustness. In columns (3) to (5), the median and mean values of ψ_k are presented, derived from the Instrumental Variable (IV) regressions of the PEESE, PET-PEESE, and EK models. These values are accompanied by 95% confidence intervals (CIs), which are constructed using *t*-statistics for the mean and bootstrapping with multiple repetitions for the median. The data set has been winsorized at the 1st and 99th percentiles. The number of meta-studies included in this analysis has been reduced to 206, as psi_k values of regressions with first-stage F-statistics less than 10 have been excluded. The data set comprises estimates exclusively from published papers.

5 Conclusion

In this study, I have conducted an analysis of a comprehensive meta-dataset comprising more than 200,000 estimates from more than 19,000 studies across 400 different fields. Utilizing key meta-regression methodologies, I present substantial evidence of selective reporting of coefficient estimates within studies that also find their way into the published literature.

This paper highlights the importance of p-hacking in the academic literature, contributing to the emerging body of work such as Brodeur et al. (2023), Lang (2023), Irsova, Doucouliagos, et al. (2023). It supports the issues raised by Irsova, Bom, et al. (2023), underscoring the critical need for meta-analytical methodologies that address the biases of p-hacking in conjunction with selection biases across studies. Furthermore, the paper underscores the risks posed by practices such as p-hacking and method searching to the robustness of established academic beliefs. It provides evidence challenging the notion that these practices are merely concerns for unpublished research, indicating their broader implications in the field.

References

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766–94.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Ashenfelter, O., Harmon, C., & Oosterbeek, H. (1999). A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour economics*, *6*(4), 453–470.

Balima, H. W., Kilama, E. G., & Tapsoba, R. (2020). Inflation targeting: Genuine effects or publication selection bias? *European Economic Review*, *128*, 103520.

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation research*, *116*(1), 116–126.

Bom, P. R., & Rachinger, H. (2019). A kinked meta-regression model for publication bias correction. *Research synthesis methods*, *10*(4), 497–514.

Brodeur, A., Carrell, S., Figlio, D., & Lusher, L. (2023). Unpacking p-hacking and publication bias. *American Economic Review*, *113*(11), 2974–3002.

Brodeur, A., Cook, N., & Heyes, A. (2020). Methods matter: P-hacking and publication bias in causal analysis in economics. *American Economic Review*, *110*(11), 3634–3660.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1–32.

Bruns, S. B., Asanov, I., Bode, R., Dunger, M., Funk, C., Hassan, S. M., Hauschildt, J., Heinisch, D., Kempa, K., König, J., et al. (2019). Reporting errors and biases in published empirical findings: Evidence from innovation research. *Research Policy*, *48*(9), 103796.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature human behaviour*, *2*(9), 637–644.

De Long, J. B., & Lang, K. (1992). Are all economic hypotheses false? *Journal of Political Economy*, *100*(6), 1257–1272.

Doucouliagos, C., & Stanley, T. D. (2013). Are all economic facts greatly exaggerated? theory competition and selectivity. *Journal of Economic Surveys*, *27*(2), 316–339.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629–634.

Ferraro, P. J., & Shukla, P. (2020). Feature—is a replicability crisis on the horizon for environmental and resource economics? *Review of Environmental Economics and Policy*.

Furukawa, C. (2019). Publication bias under aggregation frictions: From communication model to new correction method. *Unpublished Paper, Massachusetts Institute of Technology*.

Greene, W. H. (1990). *Econometric analysis*. Pearson.

Havránek, T. (2010). Rose effect and the euro: Is the magic gone? *Review of World Economics*, *146*(2), 241–261.

Havránek, T. (2015). Measuring intertemporal substitution: The importance of method choices and selective reporting. *Journal of the European Economic Association*, *13*(6), 1180–1204.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153–161.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 61–85.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, *7*(2), 246–255.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124.

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The Power of Bias in Economics Research. *The Economic Journal*, *127*(605), F236–F265.

Irsova, Z., Bom, P. R., Havranek, T., & Rachinger, H. (2023). Spurious precision in meta-analysis.

Irsova, Z., Doucouliagos, H., Havranek, T., & Stanley, T. (2023). Meta-analysis of social science research: A practitioner's guide.

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, 109–117.

Jackson, C. K., & Mackevicius, C. L. (2023). What impacts can we expect from school spending policy? evidence from evaluations in the us. *American Economic Journal: Applied Economics*.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. 289). John wiley & sons.

Kranz, S., & Pütz, P. (2021). Rounding and other pitfalls in meta-studies on p-hacking and publication bias: A comment on brodeur et al.(2020). *Available at SSRN 3848786*.

Lang, K. (2023). *How credible is the credibility revolution?* (Tech. rep.). National Bureau of Economic Research.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*(1), 31–43.

Mathur, M. (2022). Sensitivity analysis for p-hacking in meta-analyses. *OSF preprints*.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., et al. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31.

Stanley, T. D. (2005). Beyond publication bias. *Journal of economic surveys*, *19*(3), 309–345.

Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and statistics*, *70*(1), 103–127.

Stanley, T. D., Doucouliagos, H., et al. (2007). Identifying and correcting publication selection bias in the efficiency-wage literature: Heckman meta-regression. *Economics Series*, *11*, 2007.

Stanley, T. D., & Doucouliagos, H. (2012). *Meta-regression analysis in economics and business*. routledge.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78.

Van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological methods*, *20*(3), 293.

van Aert, R. C., & Van Assen, M. (2021). Correcting for publication bias in a meta-analysis with the p-uniform* method. *Manuscript submitted for publication Retrieved from: https://osfio/preprints/bitss/zqjr92018*.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*, 419–435.

Wooldridge, J. M. (2002). Econometric analysis of crosssection and panel data.