# Disentangling P-hacking From Publication Bias

Nino Buliskeria[1]

[1] *Institute of Economic Studies, Charles University; Prague, Czech Republic.*

November 19, 2023

Click here for the latest version

**Abstract**

This study differentiates between p-hacking and publication bias by identifying biases related to selective coefficient reporting as Selection Within Studies (SWS) and biases arising from selective publication as Selection Across Studies (SAS). Analyzing correlations between point estimates and their standard errors in a dataset of 400 meta-studies, which includes nearly 200,000 estimates from about 19,000 individual studies in economics and related social sciences, the research finds a consistently higher prevalence of SWS compared to SAS. This result highlights the significant impact of practices like p-hacking and method searching on selection bias in the economic literature, potentially skewing the perceived robustness of published results. The study highlights the need to simultaneously address the biases resulting from p-hacking and cross-study selection, extending the traditional concept of publication bias beyond just Selection Across Studies (SAS).

**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. I am thankful to Tomas Havranek, Jaromir Baxa, and Ali Elminejad for their helpful comments and suggestions. The responsibility for all remaining errors and omissions rests solely on me.

# 1 Introduction

Selective reporting of empirical results may distort our understanding of how robust the documented regularities are and give a false impression of their generalizability. In their influential meta-analysis, Card and Krueger (1995) addressed the pivotal question: Does raising the minimum wage reduce employment? Challenging standard economic theory, their findings famously indicated that studies corroborating a negative correlation between higher minimum wages and job availability were potentially compromised by specification-searching and publication biases. This meta-study was part of a long-term research effort for which David Card won the 2021 Nobel Prize in economics.

From the beginning of the 1980s, the critical examination of empirical research initiated by Edward Leamer catalyzed what is now broadly known as the credibility revolution in economics, which has placed a strong emphasis on meta-research and the importance of replicability of published work. This wave of change has influenced research beyond the economics, impacting fields such as medicine and epidemiology with John P. A. Ioannidis at the forefront (Begley & Ioannidis, 2015; Ioannidis, 2005; Ioannidis et al., 2017), psychology and behavioral economics to address what is commonly referred to as the "replication crisis" (Camerer et al., 2018). An expanding body of work explores the issues of potential publication biases and specification search 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).

Statistical techniques for detecting and adjusting for publication bias can be broadly categorized into two main groups. The first group consists of traditional methods derived from funnel plot analysis and the Greene (1990) "incidental" truncation theorem (Bom & Rachinger, 2019; Egger et al., 1997; Furukawa, 2019; Ioannidis et al., 2017; Stanley, 2008; Stanley & Doucouliagos, 2014) Duval & Tweedie 2000, Stanley & Doucouliagos 2012, Duval and Tweedie, 2000; Egger and Smith, 1997). This strand of the literature assumes that coefficient estimates that are statistically significant in a desirable direction are more

likely to be published (Stanley and Doucouliagos, 2014). The second group focuses on modeling the relationship between *the probability of a study's publication* and its *p-value*. These models define a parametric structure for the distribution of *population effects* before selection (Andrews & Kasy, 2019; Hedges, 1984, 1992; Iyengar & Greenhouse, 1988; Van Assen et al., 2015; van Aert & Van Assen, 2021; Vevea & Hedges, 1995). For example, in two-parameter selection models, the selection function might be designed to favor the publication of affirmative results (positive point estimates with a p-value $< 0.05$) over non-affirmative results (negative point estimates with a p-value of 0.05 or higher). By applying inverse probability weighting with maximum likelihood estimation to each study's contribution, they can jointly estimate the meta-analytic mean and the selection function's parameters.

These techniques generally conceptualize publication bias as a filtering mechanism that affects a set of point estimates that are, on their own, unbiased estimators of the true population effects (Mathur, 2022). Traditionally viewed, publication bias acts as a sieve through which studies are assessed, encompassing the choices made by researchers to refrain from submitting their study for publication, as well as the subsequent decisions by journal editors and peer reviewers whether to publish. Mathur (2022) referes to this kind of bias, resulting from various levels of selection through the research and publication process, as "selection across studies" (SAS).

However, within individual studies, the results are often vulnerable to manipulation or selective reporting, a practice known as "specification search," "p-hacking," or "data dredging" (Brodeur et al., 2022; Brodeur et al., 2020; Lang, 2023; Mathur, 2022). Actively seeking specifications that yield significant results can alter both the effect size and the standard error, leading to artificially precise results, or "spurious precision" (Irsova, Doucouliagos, et al., 2023). This presence of spurious relationships undermines a fundamental assumption of meta-analysis, particularly in selection models and regression analyses. The reliability of these methods hinges on the unbiasedness of the point estimates and their standard errors. If this condition is not met, it significantly undermines the trustworthiness of the results derived from these methodologies. Although theoreti-

cally the difference between publication bias and p-hacking is distinct, in the literature they are observationally equivalent.

In this paper, I distinguish between Selection Across Studies (SAS) and Selection Within Studies (SWS) in publication bias by analyzing how the correlation between point estimates and their standard errors vary within and across studies. The study uses a data set comprising approximately 400 meta-studies, which includes nearly 200,000 estimates derived from around 19,000 individual studies. Each meta-study pertains to specific topics within economics and related social science disciplines. The analysis employs two regression methods for each meta-analysis: Between-Effect Regression and Fixed-Effect (or Within-Effect) Regression. I use this dual-regression approach to determine the extent of bias present in both analyses by computing and comparing their respective ratios. The results demonstrate a consistently higher level of bias in fixed-effect analyzes compared to between-effect analyzes. 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.

I concentrate on five key bias correction estimators: the Egger equation (also known as the precision effect test, PET), quantile regression, the precision effect estimate with standard errors (PEESE), the combined PET-PEESE approach, and the Endogenous Kink (EK) model. My primary objective is to evaluate the extent of selection bias that arises from within-study manipulations (p-hacking, method searching) as opposed to across-study biases (biased selection for publication and the file drawer effect). For this analysis, I adopt the instrumental approach detailed by Irsova, Bom, et al. (2023).

The structure of this paper is as follows. Section 2 delves into the theoretical foundations of the bias detection techniques employed in this study. Section 3 examines the data in detail. In Section 4, I introduce the empirical techniques and discuss the results derived from these methods. The paper concludes with a final section that summarizes the study's findings and implications.

## 2 Theoretical Framework

According to the traditional definition of publication bias, each result is selected for publication based on its direction and statistical significance. This selective process is often described as an 'incidental' truncation, a term used by Wooldridge (2002). The rationale is that some coefficients are unobserved due to the results of their respective significance level. Therefore, while the objective is to determine the true mean of the coefficient estimates of interest, what is observable in the literature is a truncated distribution of these estimates. Greene (1990) presents a theorem for the moments of a truncated normal distribution. According to this theorem, if $x$ is normally distributed with mean $\mu$ and variance $\sigma^2$ , and there is a constant truncation point $a$, then:

$$E[y|truncation] = \mu + \sigma\lambda(\alpha) \tag{1}$$

where $\alpha = (a - \mu)/\sigma$, $\phi(\alpha)$ is the standard normal density and

$$\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)] \qquad \text{if truncation is } y > a,$$
$$\lambda(\alpha) = -\phi(\alpha)/\Phi(\alpha) \qquad \text{if truncation is } y < a,$$

Thus, $\lambda(\alpha)$ represents the inverse Mills ratio, which is the ratio of the probability density function to the complementary cumulative distribution function. The term $\mu$ signifies the 'true' effect, which in the context of meta-analysis would be the expected mean of the population distribution, while $\sigma$ denotes the standard error. Assuming that $\mu = \boldsymbol{x}'\boldsymbol{\beta}$ from the deterministic part of the classical regression model $y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon$ satisfies the classical assumptions on error term and dependent variable, $y$: $\epsilon|\boldsymbol{x}\ N(,\sigma^2)$ and $y|\boldsymbol{x}\ N(\boldsymbol{x}'\boldsymbol{\beta},\sigma^2)$. Greene (1990) shows that the expected value of $y$ given that it is truncated above the truncation point $a$, is as follows:

$$E[y|z > a] = \boldsymbol{x}'\boldsymbol{\beta} + \sigma\frac{\phi[(a - \boldsymbol{x}'\boldsymbol{\beta})/\sigma]}{1 - \Phi[(a - \boldsymbol{x}'\boldsymbol{\beta})/\sigma]} \tag{2}$$

Therefore, conditional mean is somewhat complex nonlinear function of $a$, $\sigma$, $x$, and

$\beta$, and hence, unfortunately, the second term of the equation, $\lambda(\alpha)$, is not constant with respect to $\mu$ and $\sigma_i$. To express the complexity of this term, I take derivative of $E[y|truncation]$ with respect to $\sigma$:

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

A common approach in the literature for detecting bias is to employ a truncated regression model. This model aims not only to determine the presence of bias, but also, ideally, to uncover the mean the mean of the target coefficient, adjusted for bias. Egger's equation is the workhorse of the meta-literature, it is a linear approximation to model the relationship between the reported effect size and its standard error to test for the existence of publication bias.

$$E_i = \gamma_1 + \alpha_1 SE_i + \epsilon \tag{3}$$

$$t_i = \alpha_1 + \gamma_1(1/SE_i) + u_i \tag{4}$$

where $\alpha = (a - \gamma_1)/SE_i$, $\phi(\alpha)$ is the standard normal density and

$$\begin{aligned} SE_i\alpha_1 &= SE_i\phi(\alpha)/[1 - \Phi(\alpha)] && \text{if truncation is } SE_i > a, \\ SE_i\alpha_1 &= -SE_i\phi(\alpha)/\Phi(\alpha) && \text{if truncation is } SE_i < a, \end{aligned}$$

where $\gamma_1$ is the mean without bias, $\alpha_1$ measures the extent of bias which is a complex function of SE, true mean, and p-value, $a$, at which observed distribution is truncated. $SE$ stands for the standard error of the specific coefficient estimate. To alleviate heteroskedasticity, the second equation is estimating Egger's equation using weighted least squares, weighted by precision, where $t_i$ is the commonly reported t-value. The test $H_0 : \gamma_1 = 0$ is known as the *precision effect test* (PET) in the literature and provides a valid test to determine whether there is a non-zero empirical effect after correcting for publication bias (Stanley, 2008); however, Stanley, 2014 shows that $\hat{\gamma}_1$ is downward

biased.

The test introduced by Egger et al. (1997) is connected to the symmetry of the corresponding funnel graph. A funnel graph, which plots the precision ($1/SE_i$) against the effect$_i$, is widely used in systematic reviews as a visual tool to indicate the presence of publication selection, as discussed in studies by Egger et al. (1997), Duval and Tweedie (2000), Hopewell et al. (2009). To alleviate the effect of heteroscesasticity, this equation is always estimated with weighted least squares (the weights being precision $1/SE$ or squared of precision $1/SE^2$).

As noted by Stanley and Doucouliagos (2014), Egger's equation can correctly measure the extent of bias and identify the mean beyond bias if the underlying empirical effect is zero ($\mu = 0$) guaranteeing a constant $\lambda(\alpha)$ and a linear relation between the expected effect and standard error. It is also noteworthy that while Egger's equation struggles to correctly identify the true mean $\mu$ beyond publication bias, it is still a valid test for the existence of publication bias. As shown in Andrews and Kasy (2019) and also intuitively from the description of the theoretical foundation, in the absence of publication bias, that is, if the probability of publication is constant with respect to the standard error, it results in the absence of truncation, and equation (1) would read $E^*[y] = \boldsymbol{x'\beta}$ given that $E^*$ denotes the best linear predictor. Therefore, testing that $H_0 : \gamma_1 = 0$ provides a valid test for the null hypothesis of no selectivity. Therefore, while not perfect, Equations (3) and (4) have been workhorse bias detection techniques of Egger et al. (1997) cited 45 196 times; Stanley (2008); Stanley and Doucouliagos (2014) and have been widely applied through different disciplines. Card and Krueger were one of the firsts to use Egger's regression in their 1995 meta-analysis, which appeared in the American Economic Review. Their study, which examined the impact of minimum wage on employment, was part of their long-term research endeavor, which eventually brought David Card the Nobel Prize in Economics in 2021.

Stanley and Doucouliagos (2014) offer an approximation of the publication bias term, $\lambda(\alpha)SE$, using alternative methods. He points out that conditional mean is nonlinear

function of $\sigma_i$ and offers emplying Taylor approximation and power series:

$$\text{effect}_i = \beta_1 + \sum_{k=1}^{K} \alpha_k SE_i^k + \epsilon_i \tag{5}$$

According to Stanley and Doucouliagos (2014), estimates of $\beta_1$ obtained from the Taylor polynomial approximation detailed in Equation (4) act as proxies for the *true* effect, $\mu$, adjusted for publication bias. Therefore, Stanley and Doucouliagos (2014) recommends adopting a quadratic approximation approach, specifically using the weighted least squares (WLS) estimate of the parameter $\beta_1$.

$$\text{effect}_i = \beta_1 + \alpha_2 SE_i^2 + \epsilon_i \quad \textbf{or} \tag{6}$$

$$t_i = \alpha_2 SE_i + \beta_1 (1/SE_i) + u_i \tag{7}$$

where meta-regression (6) is using $1/SE$ or $1/SE^2$ as the weights for the weighted least squared estimation. In the literature, $\hat{\beta}_1$ is called *precision effect estimate with standard error* (PEESE) (Havránek, 2010; Stanley, Doucouliagos, et al., 2007; Stanley & Doucouliagos, 2012).

The Precision Effect Estimate with Standard Error (PEESE) approach outperforms the linear approximation $\gamma_1$ in cases where there is a significant nonzero effect. Conversely, when there is no substantial genuine effect, the linear approximation tends to be more accurate. This pattern indicates that a combined estimator might be more effective than PEESE or $\gamma_1$ used independently. Stanley and Doucouliagos (2014) suggest employing the PEESE estimator only when there is evidence of a nonzero effect (i.e., rejecting $H_0 : \gamma_1 = 0$) as indicated in Equations (3) and (4). On the contrary, when the precision effect test (PET) is not significant (accepting H0: $\gamma_1$=0), $\gamma_1$ should be adopted as the adjusted estimate. This approach of using a conditional estimator is called 'PET-PEESE'.

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.

$$SE_i\alpha_1 = SE_i\phi(\alpha)/[1-\Phi(\alpha)] \qquad \text{if } SE_i > a,$$
$$SE_i\alpha_1 = 0 \qquad \text{if } SE_i \leq a,$$

This model employs a piece-wise linear meta-regression, where primary estimates are regressed on their standard errors.

$$effect_i = \gamma_1 + \delta[SE_i - a]I_{SE_i \geq a} + \epsilon_i \tag{8}$$

where, $\mathrm{I}_{SE_i \geq a}$ is an indicator function that takes the value of one if $SE_i$ is greater than or equal to $a$, and zero otherwise. According to Equation (8), the expected value of effect$_i$ is simply $\gamma_1 when$ $SE_i$is less than $a$, and $\gamma_1 + \delta[SE_i - a]$ when $SE_i$ is greater than or equal to $a$. Similarly to PET, PET-PEESE, the EK model addresses the heteroskedasticity of effect$_i$ by dividing each term by $1/SE_i$. The EK model endogenously determines the cutoff value based on a preliminary estimate of the true effect and a predefined threshold of statistical significance.

Using Monte Carlo simulations, Bom and Rachinger (2019) showed that the EK model is less prone to bias and more efficient compared to other regression-based methods to correct publication bias in various research scenarios. However, a limitation of the EK method is its reliance on determining a cut-off value, which can be somewhat arbitrary. This may result in reduced efficiency, especially in cases with small metasamples or low incidences of publication selection. Despite this drawback, the accuracy improvements offered by the EK method, combined with its straightforward application, make it a valuable tool for use in meta-analysis.

The implicit assumption, common in mainstream meta-regressions, including those mentioned in this section, is that the reported precision in studies accurately reflects the true underlying precision. The standard error, measured by the inverse of precision, is considered to be a measure of the true underlying precision. It is assumed that the standard errors determined by the data and methods used by the researcher are not

subject to manipulation consciously or not. (Brodeur et al., 2022; Brodeur et al., 2016; Irsova, Bom, et al., 2023; Mathur, 2022) discusses this implicit assumption and its effect on meta-analysis at length. They point out that in observational research the derivation of 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.

The methodological recommendation of Irsova, Bom, et al. (2023) is to replace the standard error reported with the portion of the error that can be explained by the sample size. They offer the Instrumental Variable approach for the Meta-analysis Instrumental Variable Estimator (MAIVE).

$$SE(\text{coeff})_i^2 = \phi_0 + \phi_1(1/N) + \nu_i \tag{9}$$

$$SE(\text{coeff})_i = \sqrt{\phi_0 + \phi_1(1/N) + \nu_i} \tag{10}$$

where Equation (9) is the first stage regression for the PEESE and PET-PEESE and Equation (10) for the PET estimation techniques; coeff is the effect size as reported in a primary study; $\psi_o$ is the constant term, $N_i$ denotes the sample size of the primary study, and $\nu_i$ is an error term. The error term of the first stage regression, $\nu_i$, absorbs the spurious components of the reported standard error that are attributable to p-hacking. Because in most contexts, the sample size is more difficult to increase than the standard error of p-hack, Irsova, Bom, et al. (2023) show that the adjusted measure is likely to better capture the underlying precision.

Irsova, Bom, et al. (2023) simulate realistic p-hacking scenario suggests that the MAIVE version of PET-PEESE, without additional inverse-variance weights, is more resistant to spurious precision than other existing methods. In situations where spurious precision plays a significant role, this method distinctly outperforms traditional unadjusted estimators, including those based on selection models. Whereas, when spurious precision is minimal, MAIVE's performance is comparable to that of unadjusted methods, though it may be occasionally outperformed by selection models, particularly in

cases with a true effect of zero. Surprisingly, they also show that the simple unweighted mean can often surpass more complex unadjusted estimators when the ratio of standard error (SE)-selection is very low relative to the total amount of selection in the simulation. Irsova, Bom, et al. (2023) conclude that solutions designed to address publication bias can sometimes create more problems than they solve.

In this paper, my focus is on the five bias-correction estimators mentioned above: linear meta-regression with study-level fixed effects and between-effects meta-regression, quantile regression, Precision-Effect Estimate with Standard Errors (PEESE), PET-PEESE, and the Endogenous Kink (EK) model. The primary aim 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). However, for the sake of developing intuition and maintaining simplicity, I begin with Egger's 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 and Methodology

## 3.1 Data description

This thesis investigates the sources of selective reporting by examining within study selection and across study selection in 400 meta-analyzes, which encompasses more than 20,000 studies and 200,000 coefficient estimates from various fields of social sciences, mainly economics. The data set was provided in part by Chris Doucouliagos and was collected in part from previous and newly published meta-studies. It contains information on the authors, titles, publication year, and journal of meta-study, as well as for the studies. Furthermore, the metadata contains coefficient estimates, their respective standard errors, and the sample size from each study.

Many meta-studies examine questions that are closely related, 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 range of macroeconomic indicators, including the effects of inflation targeting on inflation, GDP, interest rate volatility, inflation volatility, growth volatility, exchange rate volatility, and deficit. In my analysis, 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. Similarly to Doucouliagos and Stanley (2013), I find "substential" selectivity across 400 different topics and "sevier" in under 100 topics in economics & social sciences (see Figure 6).

Figure 1 presents a histogram of the number of meta-studies published each year within the data set. This trend aligns with the overall growth in meta-study publications in the literature, surpassing 100,000 publications by 2022 (Irsova, Doucouliagos, et al., 2023). The primary data set compiled for this article consists predominantly of studies published before 2021. However, data sets from metastudies published in 2021 and 2022 were subsequently incorporated to ensure the relevance of the data set.

An analysis of the journals where these meta-studies have been published reveals a concentration in a wide array of economic disciplines. Figure 2 provides a visual representation of this distribution, categorizing research areas according to the SCImago Journal Rank (SJR). It also shows the frequency of publications within each research area. Notably, the fields of *Economics, Econometrics and Finance*, with over 100 meta-analyses, is also mentioned as part of majority of other area classifications. The repeated appearance of the *Economics, Econometrics, and Finance* classification throughout Figure 2 indicates that our data set primarily comprises estimates drawn from economic research.

Figure 3 shows the journals that are the most frequent sources for the meta-analysis of the sample. Not surprisingly, it reflects the picture that can be seen in Figure 2, where the most recurring research area is economics. Here, in Figure 3, 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 from my sample at least twice; however, similarly to Figure 3, the economic journals are majority among all the journals, social science and interdisciplinary journals second most frequent and rarely medicine.

## 3.2 Selection Within vs. Across Study

Study-fixed-effects in meta-regression provide a straightforward way to disentangle bias-related variation into within- and between-study elements, an approach that has not been systematically exploited.

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, let us assume for now that any correlation between the coefficient $coef_{ij}$, and their standard error, $SE_{ij}$, is counterintuitive and indicates the existence of bias. Thus, the correlation between $coef_{ij}$ and $SE_{ij}$ within the study indicates bias from SWS, and the correlation between studies indicates bias due to SAS.

I perform 400 fixed effect regressions to assess within-study selection and between-effect regressions to control for between-study selection, for each meta-analysis $k$, study $j$, and estimate $i$, I have the following:

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

Where $coef_{ij}$ is the coefficient estimate $i$ from study $j$; $SE_{ij}$ is the corresponding standard error; $e_j$ indicates study-specific unobserved characteristics and $u_{ij}$ is the error term.

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

The fixed effect estimator takes care of the fixed effect of $e_j$ the unobserved study by subtracting the study mean estimates; thus, eliminating variation between studies, it studies within-study variation.

In comparison, I study between study variations using an estimator between studies

that takes averages over studies:

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

Finally, I calculate $\beta_k^{FE}$ and $\beta_k^{BE}$ and derive $\psi_k = \frac{\beta_k^{FE}}{\beta_k^{BE}}$ for each meta-study $k$ .

I estimate the $\psi_k$ ratio from linear fixed effect and between effect models, 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. In this table, I present the median and mean values of $\psi_k$ with the 95% confidence interval (CI) constructed using t statistics for mean and bootstrapping with a sample with multiple repetitions for median.

Table 1: Selection Within vs. Across Study

| | **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 $\psi_k$ are detailed, each accompanied by a 95% confidence interval (CI). These intervals are calculated using t statistics for the mean and using bootstrapping with multiple repetitions for the median. Additionally, the dataset has undergone winsorization at the 1st and 99th percentiles to enhance its statistical robustness.

Next, to alleviate the effect of outliers, I imply median regression, quantile regression at 50%, on the original data without winorization. Next, in Table 2, I show the analysis based on PEESE, PET-PEESE, and EK regressions. To control for possible p hacking and avoid overestimation of bias, I employ suggestions Irsova, Bom, et al. (2023) and use inverse of sample size to instrument for the standard errors.

In all five approaches (Tables 1 & 2), I find that the bias arising from the variation within the study is greater than the selection between studies. Although the mean value is greater than 5 in all cases, this estimate can be influenced by how scattered the $\psi_k$ values are, since we are looking at different questions and fields. Therefore, it is essential to look

Table 2: Selection Within vs. Across Study

| | **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 $\psi_k$ 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 winorized at the 1st and 99th percentiles. The number of meta-studies included in this analysis has been reduced to 206, as $psi_k$ values from regressions with first stage F-statistics less than 10 have been excluded.

at the median value of $\psi_k$. Together, the median and mean values of the ratio suggest that SWS is consistently larger compared to SAS, pointing to the prevelant evidence of practices like method searching and p-hacking in the published and working literature.

These conclusions are drawn from looking at the complete data. Next, I look at only published work to evaluate the comparison of SWS and SAS in published literature.

Table 3: Selection Within vs. Across Study, subset of published papers

| | **Linear Regression** | **Quantile Regression** |
|---|---|---|
| Median | 1.15 | 1.07 |
| Median CI | [1.03; 1.38] | [0.94; 1.45] |
| Mean | 7.37 | 6.21 |
| Mean CI | [5.07; 9.66] | [3.63; 8.79] |
| Number of Meta-Studies | 398 | 368 |

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

However, Tables 4 and 5 demonstrate that findings derived exclusively from published literature are consistent with those obtained from the entire dataset. The Selection Within Studies (SWS) is consistently found to be more pronounced than Selection Across

Table 4: Selection Within vs. Across Study, subset of published papers

| | **PEESE** | **PET-PEESE** | **EK** |
|---|---|---|---|
| Median | 1.33 | 1.29 | 1.22 |
| Median CI | [ 1.15; 1.51] | [1.05; 1.76] | [1.07; 1.44] |
| Mean | 7.44 | 7.58 | 4.41 |
| Mean CI | [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 $\psi_k$ 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 $psi_k$ values from regressions with first-stage F statistics less than 10 have been excluded. The data set comprises estimates exclusively from published papers.

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.

# 4 Conclusion

In this study, I have conducted an analysis of a comprehensive meta-dataset comprising over 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 published literature.

This paper highlights the significance of p-hacking in the academic literature, contributing to the emerging body of work by researchers such as Brodeur et al. (2022), 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 biases from 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.

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. (2022). *Unpacking p-hacking and publication bias* (tech. rep.). University of Ottawa.

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.

Card, D., & Krueger, A. B. (1995). Time-series minimum-wage studies: A meta-analysis. *The American Economic Review*, *85*(2), 238–243.

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.

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.

Hopewell, J., Dvorak, R., & Kosior, E. (2009). Plastics recycling: Challenges and opportunities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1526), 2115–2126.

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.

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.

# Figures

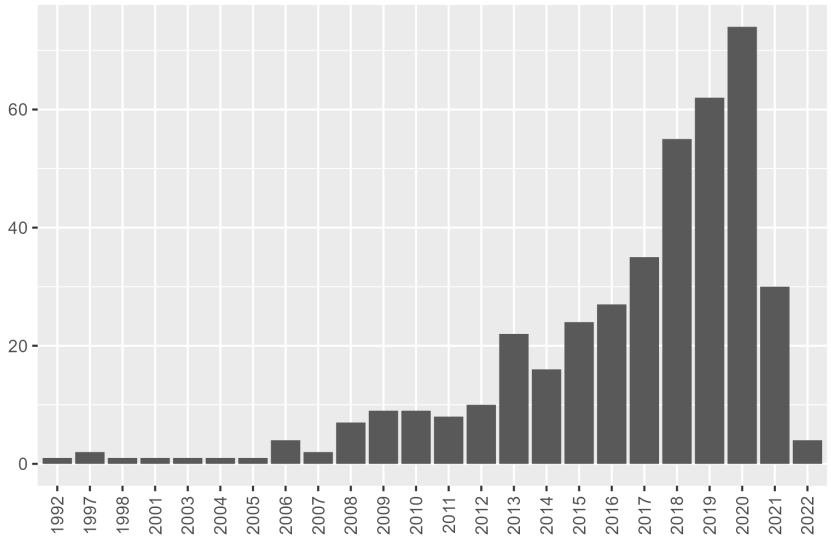


Figure 1: year meta-study was published in the jouranl or online platform as working paper.

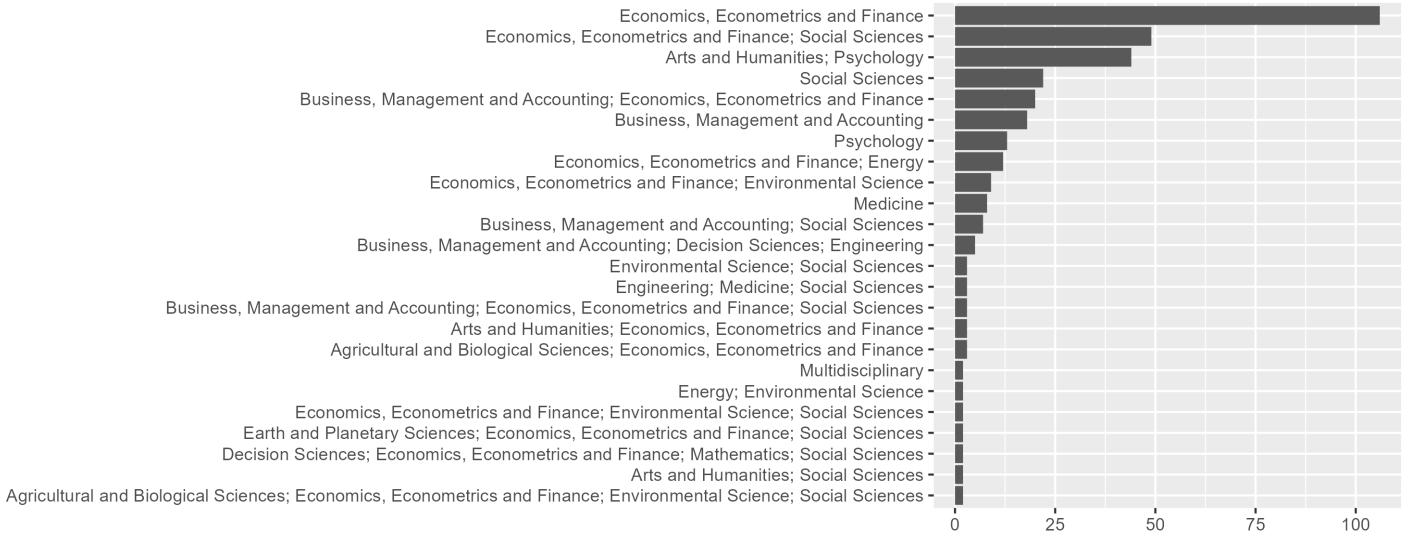


Figure 2: shows the number of meta-analysis published in journals with each research areas classification according to the SCImago Journal Rank (SJR)

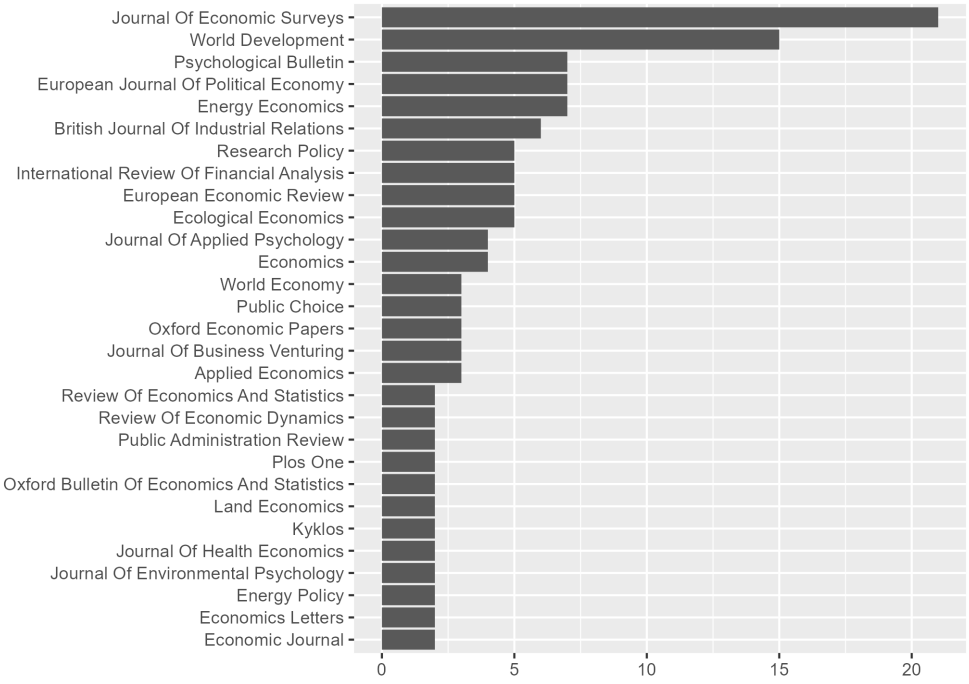


Figure 3: provides a list of journals that are the most frequent publishers of meta-studies included in the dataset.

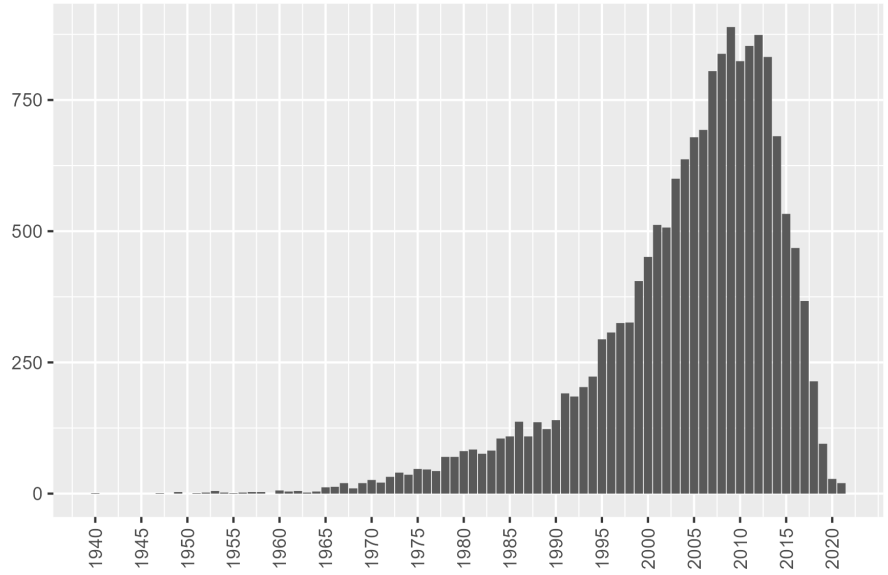


Figure 4: distribution of published studies in meta-data over years

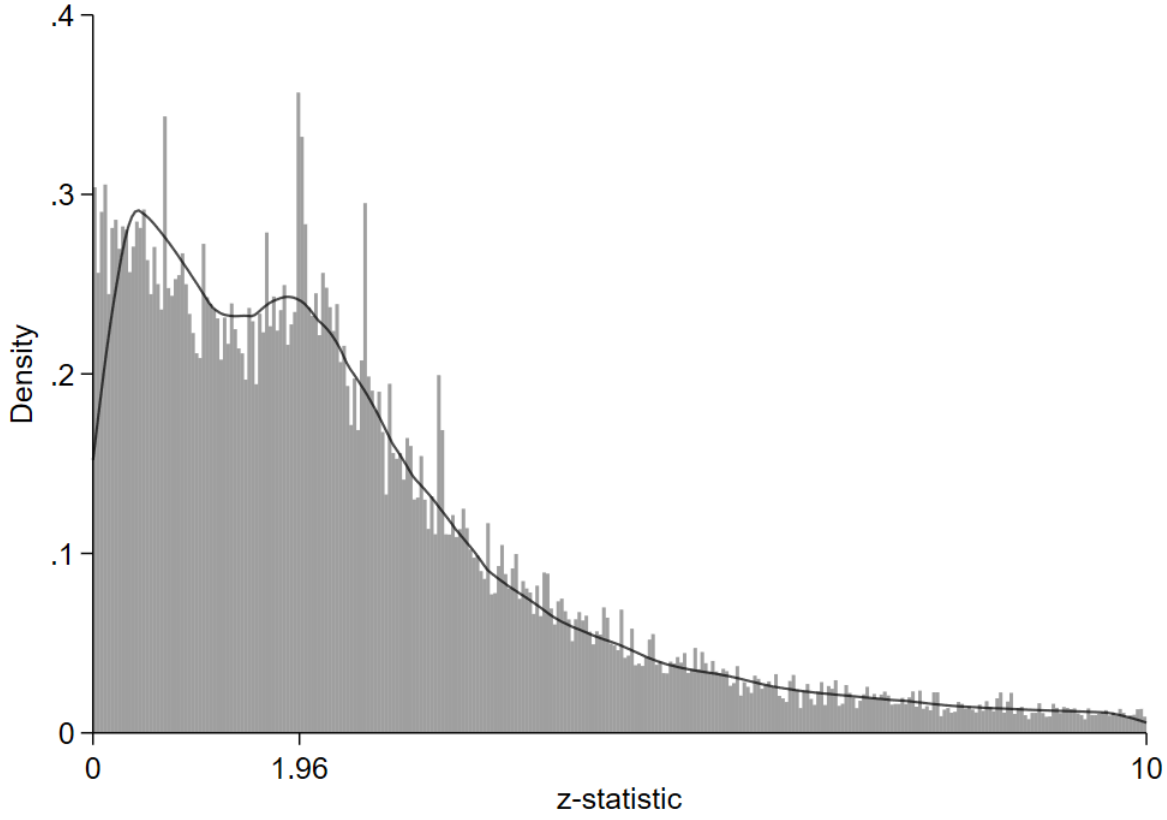


Figure 5: De-rounded & weighted distribution of z-statistics of published papers. Note: The two-humped camel-shaped pattern, similar to Brodeur et al. (2020, 2023) is evident.

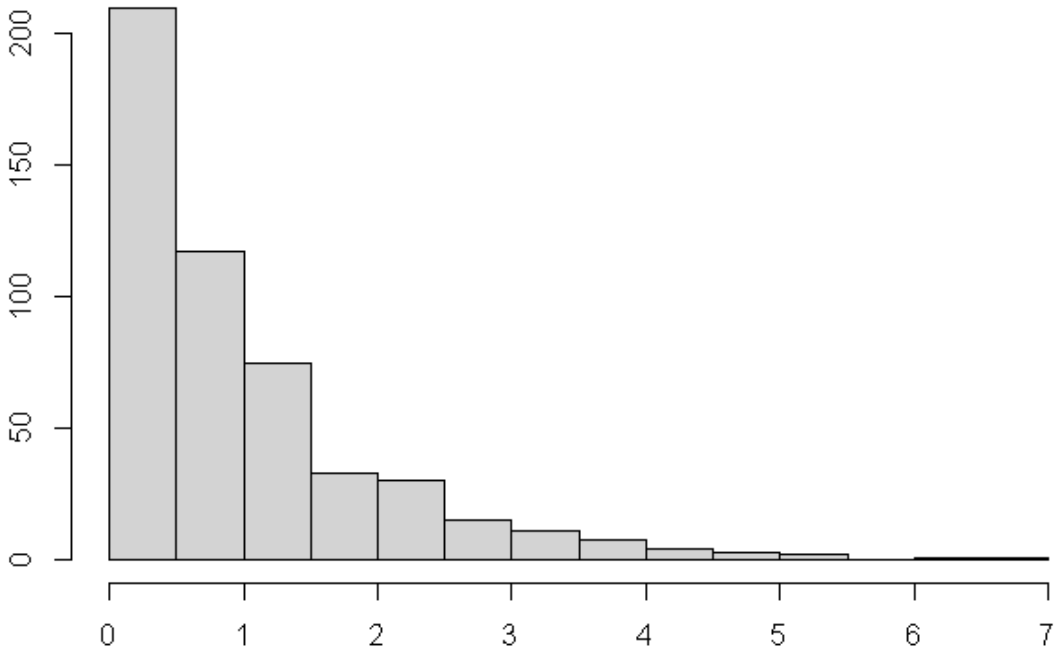


Figure 6: Distribution of Selectivity in Empirical Economics. Bias estimated from the Egger's regression, $coef_i = \alpha + \beta SE_i + \epsilon_i$. The bias is considered "littel to modest" if $|\beta > 1|$, "substential" if $1 \leq |\beta| \leq 2$ and "sevier" for $|\beta > 2|$. Similar to Doucouliagos and Stanley (2013), I find "substential" selectivity across 400 different topics and "sevier" in under 100 topics in economics & social sciences .