Disentangling P-hacking From Publication Bias

Nino Buliskeria[1]

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

November 30, 2023

Click here for the latest version

Abstract

This study distinguishes between p-hacking and publication bias by examining biases stemming from selective reporting within studies versus selective publication of entire studies. Analyzing a dataset of 400 meta-studies, encompassing 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 as compared to selective publication. Employing 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 significantly contribute 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. 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-analyzes, 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 economics to address what is commonly referred to as the "replication crisis" (Camerer et al., 2018), affecting 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. 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 to detect and adjust 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 line 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 *a study's publication probability* 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 evaluated, 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 on whether to publish. Mathur (2022) refers 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-analyses, particularly in selection models and regression analyses. The reliability of these methods depends 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

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

Although the literature agrees on the potential consequences of published phacked coefficients, the magnitude of the matter or the way to measure the extent of p-hacking in the literature is ambiguous. Although Brodeur et al. (2022) shows convincing arguments in favor of the dominant role of p-hacking in publication bias, Lang, 2023 does not find evidence to support this phenomenon. In this paper, I distinguish the effect of p-hacking from publication bias by analyzing how the correlation between point estimates and their standard errors varies within and across studies. To conduct this analysis, I use 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. I define p-hacking as the action of the authors that causes the correlation between the point estimate and the standard error within the study; otherwise, the selection bias within the study. I employ fixed-effects analysis to identify within-study selection bias, controlling for study-specific characteristics. Next, I apply different meta-regression analysis techniques on means of coefficient and standard error pairs for each study to identify the selection bias between studies, measuring the magnitude of selection across studies and the selection type that does not introduce bias in point estimates.

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 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. My analysis indicates that selective reporting of coefficients and p-hacking is 20 to 30 % larger compared to selection between studies. This result aligns with the evidence presented by Brodeur et al. (2022).

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 of these methods. The paper concludes with a final section summarizing the study's findings and implications.

2 Theoretical Foundation

According to the traditional definition of publication bias, research results are selected for publication based on their direction and statistical significance. Although this selective publication process skews the overall distribution of reported results in the literature, it is often assumed that the chosen results are unbiased estimations of the true underlying effect relative to their respective population effect. Therefore, most publication bias detection and correction techniques rely on this assumption. However, (Brodeur et al., 2022; Brodeur et al., 2016; Irsova, Bom, et al., 2023; 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 naturally cause spurious precision and violate the core assumption of unbiased estimates. Violation of this assumption renders metaregression 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 potential consequences of published phacked coefficients, the significance of the matter or the way to measure it is till ambiguous.

In this section, I discuss the theoretical foundation of meta-regression analysis (MRA) and the importance of the underlying unbiasedness assumption 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, show my identification strategy to measure the magnitude of phacking compared to selection across studies.

Similarly to Jackson and Mackevicius (2023), I start by building the discussion from the points estimates in each study. Consider a series of studies to estimate the effect size of a specific research question. Each study uses distinct sample specifications and robust techniques to achieve unbiased estimates. In this scenario, the study i produces an estimated effect, represented as $\hat{\alpha}_i$, which is expected to be close to the actual true effect, denoted as α_i. The discrepancies between these estimated and true effects result from sampling errors and measurement inaccuracies. Additionally, I follow the conventional assumption that the true effect size follows a normal distribution with a mean of Θ and a variance of $\aleph^2$:

$$\alpha_i \sim N(\Theta, \aleph^2) \tag{1}$$

Following the Central Limit Theorem[1], the distribution of the estimated effect size is:

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

This 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

[1]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, the estimates have not been standardized; hence, they are normally distributed with mean and variance.

centered around the true effect.

$$\hat{\alpha}_i \sim N(\Theta, \sigma_i^2 + \aleph^2) \tag{3}$$

Therefore:

$$\hat{\alpha}_i = \alpha_i + u_i \tag{4}$$

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

Let us now consider the classical definition of publication bias. The 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 $\hat{\alpha}_i|t_{\hat{\alpha}_i} > c$. This truncation then creates publication bias. Next, I discuss each selection type separately.

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:

$$\begin{aligned} E[\hat{\alpha}_i|\hat{\alpha}_i > a] &= \alpha_i + E[u_i|\hat{\alpha}_i > a] \\ &= \alpha_i + E[u_i|u_i > a - \alpha_i] \end{aligned} \tag{5}$$

where σ_i is estimated standard error from study i, $E[u_i|u_i > a - \alpha_i] = \sigma_i\phi(\kappa)/[1 - \Phi(\kappa)]$ 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.

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

Therefore, the meta-regression is as follows:

$$E[\hat{\alpha}_i|\hat{\alpha}_i > a] = \alpha_i + \sigma_i\lambda(\kappa) \tag{6}$$

Thus, $\lambda(\kappa)$ represents the inverse Mills ratio. If the truncation of the estimated coefficient is from above $\alpha_i|\alpha_i < a$, then $\lambda(\kappa) = -\phi(\kappa)/\Phi(\kappa)$. The term α_i signifies the 'true' effect, while $\hat{\sigma}_i$ denotes the standard error of the estimated coefficient.

The truncation of the significance is similar to the truncation of the coefficient estimate (referred to as incidental truncation in Greene (1990), Theorem 2.5 also, see Heckman (1979)). 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 $c = 1.96$). 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,

$$\hat{\alpha}_i/\sigma_i \sim N(\alpha_i/\sigma_i, 1) \tag{7}$$

with bivariate normal joint distribution. Therefore, following Theorem 2.5 in Greene (1990)[2]

$$E[\hat{\alpha}_i|\hat{t} > c] = \alpha_i + \sigma_i\rho\frac{\phi(\kappa_{it})}{1 - \Phi(\kappa_{it})} \tag{8}$$

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 expression (7), $\rho = 1$ and $\kappa_{\hat{t}} = (c - \hat{\alpha}_i/\sigma_i)$ resulting in the same form of meta-regression as shown in Equation (6).

To estimate α_i, often referred to as mean beyond bias in the meta-literature, one needs to 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 of this term, I take the derivative of $E[\hat{\alpha}|truncation]$ with respect to σ, I drop i for simplicity, however

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

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:

$$\begin{aligned}\partial\lambda(\kappa)/\partial\kappa &= \frac{\phi'(\kappa)[1-\Phi(\kappa)] + \phi(\kappa)\Phi'(\kappa)}{[1-\Phi(\kappa)]^2} \\ &= \frac{\phi'(\kappa)[1-\Phi(\kappa)] + \phi(\kappa)^2}{[1-\Phi(\kappa)]^2} \\ &= -\frac{\phi(\kappa)\cdot\kappa}{[1-\Phi(\kappa)]} + \frac{\phi(\kappa)^2}{[1-\Phi(\kappa)]^2} \\ &= \lambda^2(\kappa) - \kappa\cdot\lambda(\kappa)\end{aligned} \tag{9}$$

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

$$\partial E[\hat{\alpha}|truncation]/\partial\sigma = \lambda(\kappa) + \frac{\alpha}{\sigma}[\lambda^2(\kappa) - \kappa\cdot\lambda(\kappa)] \tag{10}$$

Equation (6) is the statistical foundation of the meta-regression model for bias detection, and Equation (10) shows the relation between the expected mean of truncated estimates and their standard error.

A common approach in the literature to detect bias is to employ a truncated regression model (see Equation 6), also known as the Egger's equation.[3]

$$\hat{\alpha}_i = \alpha + \lambda\sigma_i + \epsilon_i \tag{11}$$

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

[3]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 notations.

precision, where t_i is the reported t statistics.

$$t_i = \lambda + \alpha(1/\sigma_i) + u_i \tag{12}$$

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 non-zero effect size. This is intuitive once we compare Equation (11) with (6), 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 **??** & 10. 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 10 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, non-zero cases remain challenging for PET approach.

The literature strand successfully addresses this issue, using different weighting and Taylor approximation techniques to appeal to the second-order structure of the Equation 10 (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.

$$\hat{\alpha}_i = \alpha + \lambda\sigma_i^2 + \epsilon_i \qquad \text{or} \tag{13}$$

$$t_i = \lambda\sigma_i + \alpha(1/\sigma_i) + u_i \tag{14}$$

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 14 only when there is evidence of a non-zero effect (i.e.,

rejecting $H_0 : \alpha = 0$), and the PET estimator, Equation (11), when accepting $H_0 : \alpha = 0$, resulting in 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. Hence, the EK model approximates $\lambda(\kappa)$ using a piece-wise linear meta-regression:

$$\hat{\alpha}_i = \alpha + \delta[\sigma_i - a]I_{\sigma_i \geq a} + \epsilon_i \tag{15}$$

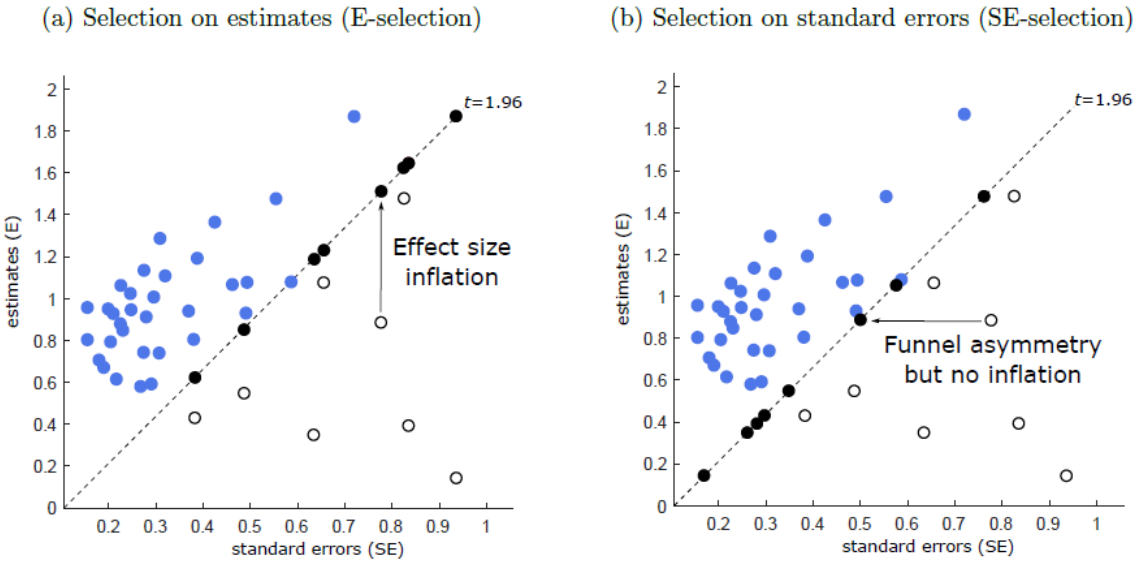
where, $\mathrm{I}_{\sigma_i \geq a}$ is an indicator function that takes the value of one if σ_i is greater than or equal to a, and zero otherwise. Similarly to PET, PET-PEESE, the EK model addresses the heteroskedasticity of $\hat{\alpha}_i$ by dividing each term by $1/\sigma_i$. The EK model endogenously 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 the 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 1, taken from Irsova, Bom, et al. (2023), illustrates the distributional consequences of various actions such as cheating, clustering, correcting for heteroskedasticity and addressing nonstationarity, all undertaken to obtain statistically significant results without a solid theoretical or reasonable basis. This figure distinguishes between the

distributional effects of selection based on estimates (panel a), which is typical in the existing literature, and selection based on standard errors (panel b). The panel (a) of

Figure 1: Spurious precision renders common meta-analysis techniques biased



Reference: Irsova, Z., Bom, P. R., Havranek, T., & Rachinger, H. (2023). Spurious precision in meta-analysis, available at meta-analysis.cz/maive.

Figure 1 shows 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 this manipulation, 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, panel (b) of Figure 1 shows the cases where researchers achieve statistical significance by reducing the standard error. However, in this case, there is no bias in the reported effect sizes; both the black-filled and the hollow circles 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 introduces an additional downward bias.

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 Meta-analysis Instrumental Variable Estimator (MAIVE) model, where they instrument standard error with the inverse of sample size. Since 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 captures the underlying precision better.

$$\sigma_i^2 = \phi_0 + \phi_1(1/n_i) + \nu_i \tag{16}$$

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

where Equation 16 is the first stage regression for the PEESE and Equation 17 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 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 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). 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. For the sake of developing intuition and maintaining simplicity, I begin with the linear 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.

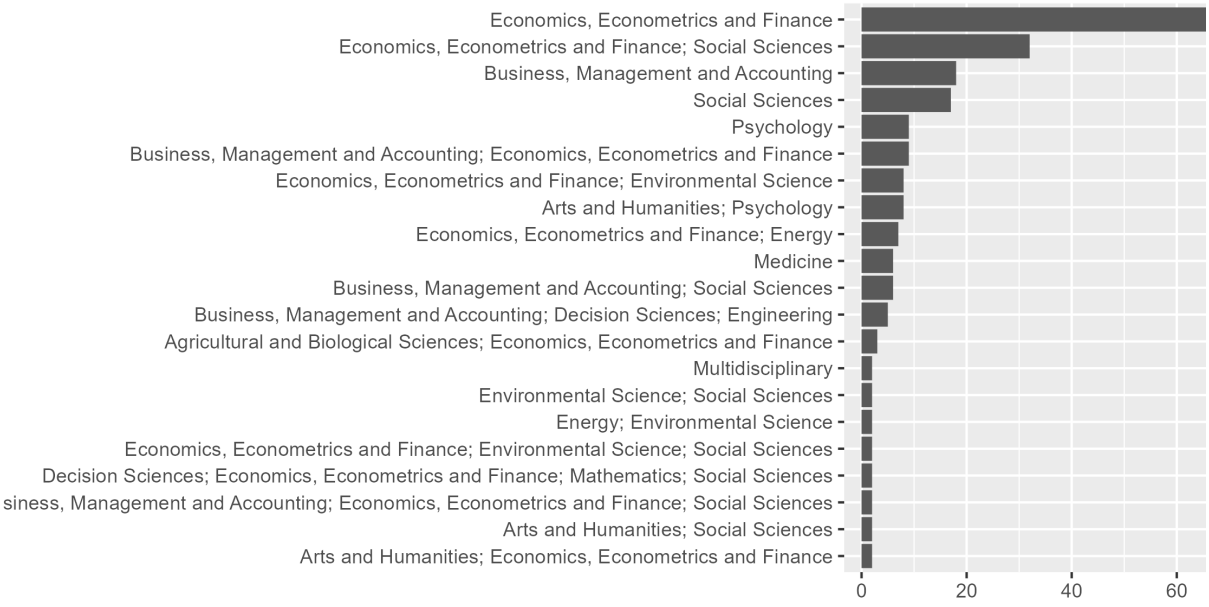
3 Data description

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. Furthermore, the metadata contains coefficient estimates, their respective standard errors, and the sample size 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) analyzes 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.

Figure 2: The meta-analyses published in journals areas

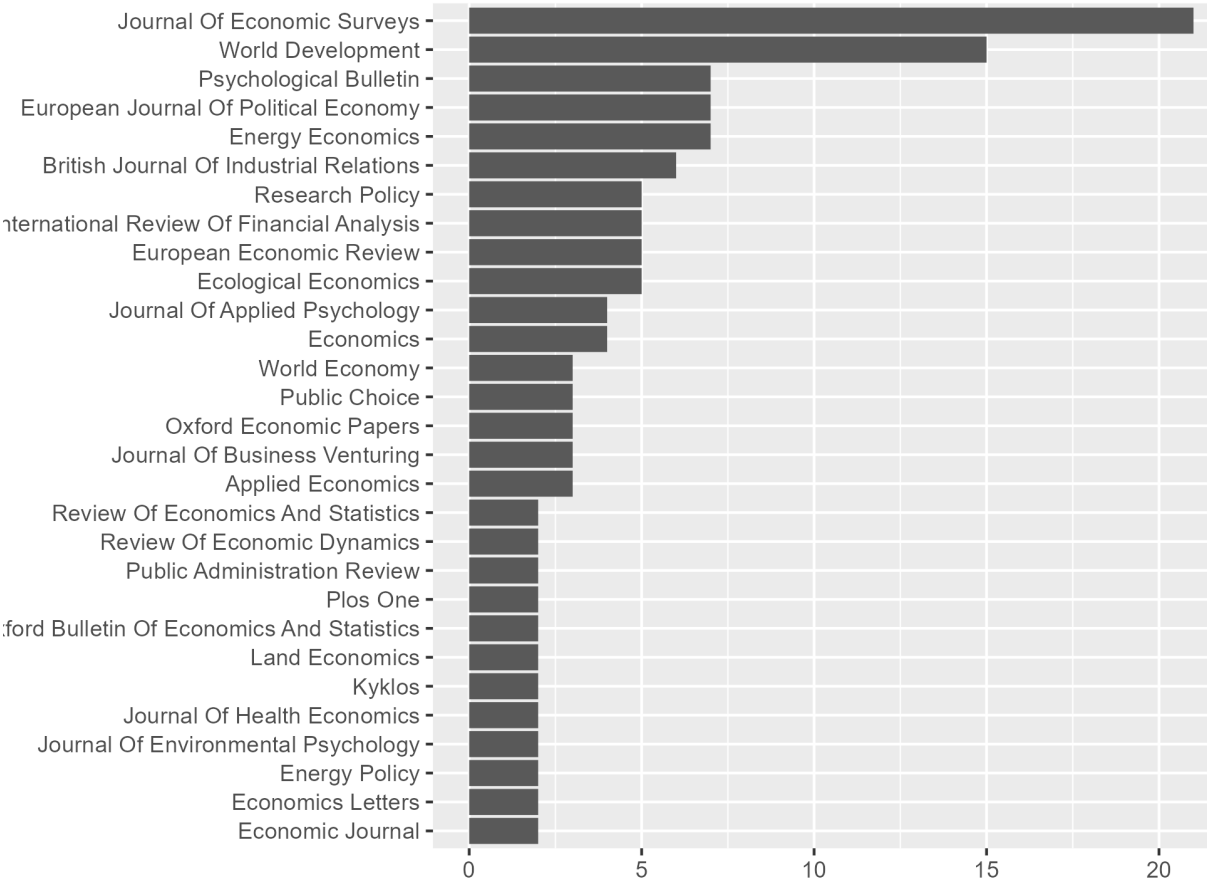


Note: Journal research areas classification according to the SCImago Science Journal Rank (SJR), https://www.scimagojr.com/journalrank.php?area=2000

An analysis of the journals where these meta-studies have been published reveals a concentration in various economic disciplines. Figure 2 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 2 indicates that our data set mainly comprises estimates drawn from economic research.

Figure 3: Meta-analyses per journal

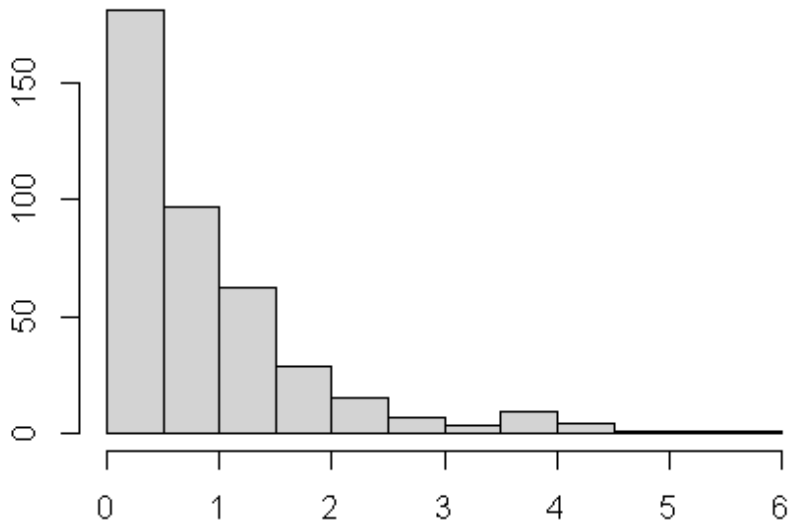


Note: a list of journals that are the most frequent publishers of meta-studies included in the dataset.

Figure 2 shows the journals that are the most frequent outlets for published meta-analyses in the data. Not surprisingly, it reflects the picture that can be seen in Figure 2, where the most frequent research area is economics. 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 in the sample at least twice; however, similarly to Figure 2, the economic journals are the majority of the journals, and social science

and interdisciplinary journals are the second most frequent and rarely medicine.

Figure 4: Distribution of Selectivity in Empirical Economics.

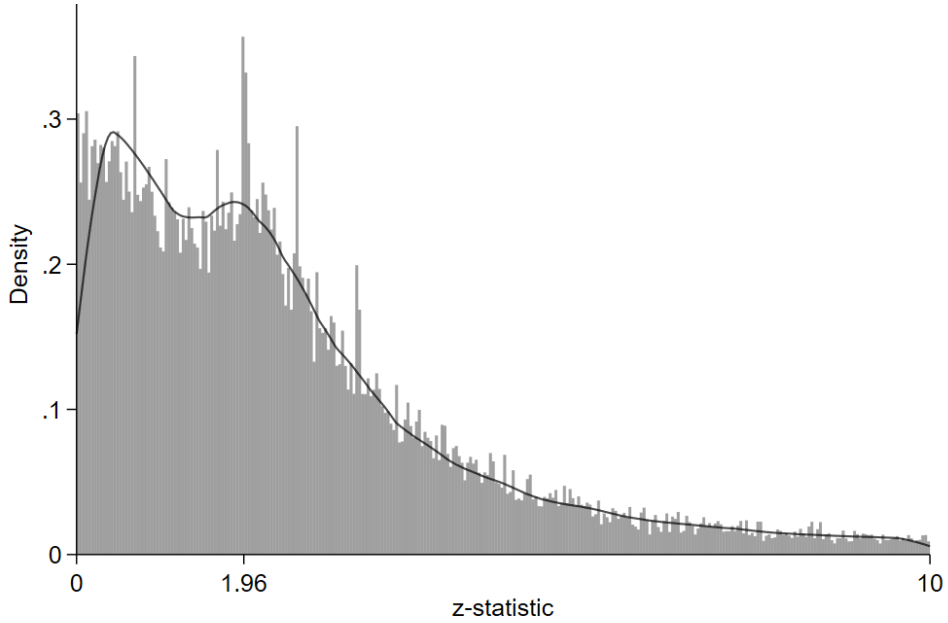


Note: Bias estimated from Egger's regression, $coef_i = \alpha + \beta SE_i + \epsilon_i$. The bias is considered *little to modest* if $|\beta < 1|$, *substantial* if $1 \leq |\beta| \leq 2$, and *severe* for $|\beta > 2|$. I find *substantial* selectivity across 91 different topics and *severe* in 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 k number of β coefficients for each topic. Figure 4 shows the distribution of β_k in 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 5, I look at the distribution of t-statistics in published articles and show evidence of potential p-hacking, as discussed in Brodeur et al. (2022). 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

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.

distribution; it only smooths potential discontinuities in histograms. Figure 5 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 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 5. 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 5. This paper

contributes to this discussion by analyzing similar questions based on metar-egression analysis.

4 Selection Within vs. Across Study

Study-fixed effects in metaregression 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 its standard error SE_{ij} 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 evaluate the selection of the within-study and between-effect regressions to control the selection of the between-study for each meta-analyses k, study j and estimate i, I have the following:

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

Where $coef_{ij}$ is the coefficient estimate i of study j; SE_{ij} is the corresponding standard error; e_j indicates characteristics specific to the study 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 for 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

taking averages over studies:

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

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

I estimate the ψ_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 ψ_k with 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 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 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 ψ_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 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 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 ψ_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 ψ_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 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 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.

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 published literature.

This paper highlights the importance 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 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.

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.

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.