Received: 15 February 2023 | Revised: 23 December 2023 | Accepted: 4 February 2024

DOI: 10.1002/jrsm.1711

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Estimating the extent of selective reporting: An application to economics

Stephan B. Bruns[1,2,3] | **Teshome K. Deressa**[1] | **T. D. Stanley**[4] | **Chris Doucouliagos**[4] | **John P. A. Ioannidis**[3,5,6,7,8]

[1] Centre for Environmental Sciences, Hasselt University, Hasselt, Belgium

[2] Department of Economics, University of Göttingen, Göttingen, Germany

[3] Meta-Research Innovation Center at Stanford (METRICS), Stanford, California, USA

[4] Department of Economics, Deakin University, Melbourne, Victoria, Australia

[5] Department of Medicine, Stanford University, Stanford, California, USA

[6] Department of Epidemiology and Population Health, Stanford University, Stanford, California, USA

[7] Department of Biomedical Data Science, Stanford University, Stanford, California, USA

[8] Department of Statistics, Stanford University, Stanford, California, USA

Correspondence

Stephan B. Bruns, Centre for Environmental Sciences, Hasselt University, Hasselt, Belgium.

Email: stephan.bruns@uhasselt.be

Funding information

German Research Foundation, Grant/Award Number: 405039391; Special Research Fund (BOF) UHasselt, Grant/Award Number: BOF20OWB05

Abstract

Using a sample of 70,399 published *p*-values from 192 meta-analyses, we empirically estimate the counterfactual distribution of *p*-values in the absence of any biases. Comparing observed *p*-values with counterfactually expected *p*-values allows us to estimate how many *p*-values are published as being statistically significant when they should have been published as non-significant. We estimate the extent of selectively reported *p*-values to range between 57.7% and 71.9% of the significant *p*-values. The counterfactual *p*-value distribution also allows us to assess shifts of *p*-values along the entire distribution of published *p*-values, revealing that particularly very small *p*-values ($p < 0.001$) are unexpectedly abundant in the published literature. Subsample analysis suggests that the extent of selective reporting is reduced in research fields that use experimental designs, analyze microeconomics research questions, and have at least some adequately powered studies.

KEYWORDS

meta-research, *p*-hacking, publication bias, selective reporting

Highlights

What is already known

- Selective reporting is prevalent in many disciplines, including economics.
- However, estimating the extent of selective reporting remains challenging.

What is new

- We estimate the extent of selective reporting in economics by comparing the distribution of observed *p*-values with the counterfactual distribution of *p*-values generated under the assumption of no biases.
- Our approach allows us to quantify the lack and abundance of observed *p*-values along the entire distribution of *p*-values.

- The observed and counterfactual distribution of p-values intersect at the 0.1 level of significance. Non-significant p-values are uniformly missing and predominantly shifted to highly significant p-values ($p < 0.001$).

Potential impact for *Research Synthesis Methods* readers

- Our work opens new avenues in appraising the credibility and selective reporting of entire scientific fields based on systematically collected data through meta-analysis.

1 | INTRODUCTION

There is increasing evidence that statistical significance might be frequently inflated in empirical economics[1–7] and other disciplines,[8–10] but quantifying the extent of selectively reported p-values is challenging. In this study, we estimate the extent of selective reporting in economics by empirically estimating a counterfactual distribution of published p-values that would have occurred in the absence of any biases in the research and publication process. We also explore potential determinants of selective reporting.

The preferential publication of statistically significant findings incentives researchers to provide such findings[11,12] and the underlying mechanism is predominantly known as selective reporting or publication selection bias,[13] such as p-hacking,[14,15] HARKing (Hypothesis After Results are Known),[16] and publication bias.[17] Selective reporting is the behavioral response of researchers who need to publish to strive for tenure, to acquire competitive research funding or to advance their career more generally.[18] The selective reporting of analyses that "work" from a potentially large set of analyses conducted in the research process was recently coined p-hacking[14] but the underlying problem has been earlier discussed under different names in economics[19,20] and statistics.[21,22] Hacking the p-value to be statistically significant is eased with researchers' degrees of freedom in the analysis being often vast in both observational research[23–26] and experimental research.[27,28]

While p-hacking describes the selection of analyses that "work" for a given hypothesis, HARKing refers to researchers that explore associations in data sets and then search for a suitable hypothesis or theory once a statistically significant finding is found.[16] Finally, p-hacking and HARKing operate at the analysis level while publication bias describes the selection of statistically significant findings at the study level.[17] Franco et al.[29] show that non-significant findings have a substantially smaller probability of being written up by authors.

It is important to emphasize that selective reporting is not necessarily conscious and intentional scientific misbehavior.[5] It may result from unconscious and "naive" experimenting with the data and researchers may be prone to motivated reasoning once a significant estimate is found.[30]

There is increasing evidence for selective reporting in economics. Ioannidis et al.[3] show that economics is largely underpowered suggesting that effect sizes are substantially overestimated to generate statistical significance. Caliper tests which compare the frequency of tests just before and after the threshold of statistical significance indicate evidence for discontinuities at the typical thresholds of statistical significance.[5,6] Moreover, numerous meta-analyses have documented the presence of selective reporting in various subfields of economics.[7,1,31]

We estimate the counterfactual distribution of published p-values that would have occurred if all studies had estimated the respective genuine effect unbiasedly. To this end, we approximate genuine effects with meta-averages, which is a conservative approach as estimated meta-averages are known to overestimate the genuine effect.[32] Comparing factually observed p-values with counterfactually expected p-values allows us to quantify how many p-values were published as being statistically significant when they should have been published as non-significant. The counterfactual p-value approach allows us to analyze the excess and lack of p-values along the entire p-value distribution.

We estimate the extent of selectively reported p-values to range between 57.7% and 71.9% of the significant p-values. We find that p-values are missing throughout the entire range of non-significant p-values while there is predominantly an excess of p-values that are below 0.001. Our subsample analyses suggest that experimental research designs, microeconomic research and research fields with at least some adequately powered studies (APS) exhibit less selective reporting.

While our findings suggest that selective reporting seems to be large in economics, many measures to improve the reliability of empirical research have been already implemented in recent years, including an emphasis on pre-registered randomized controlled trials[33,34] and a critical reflection on dichotomizing statistical findings in statistically significant and non-significant.[35] We particularly advocate in line with Ioannidis et al.[3] to routinely establish power considerations in observational research in economics to ensure that researchers have sufficiently large sample sizes to obtain statistically significant findings if the hypothesized effect exists in order to avoid significance chasing by exaggerating estimated effect sizes.

2 | EMPIRICAL STRATEGY

2.1 | Data

Our sample comprises 192 meta-analyses with a total of 70,399 coefficients with respective standard errors and is an updated version of the data used in Ioannidis et al.'s[3] study.[1] Each meta-analysis addresses a distinct research question and the included coefficients are those that try to answer this question. We identified meta-studies using search engines (Econlit, Scopus, and Google Scholar), publisher sites (e.g., Science Direct, Sage, and Wiley), and webpages of researchers known to publish meta-analyses. We also searched all volumes of individual journals that are known to publish meta-analyses, for example, *Journal of Economic Surveys*, *World Development*, *Public Choice*, *European Journal of Political Economy*, *Oxford Economic Papers*, *European Economic Review*, and *Ecological Economics*. We focus on meta-analyses published in economics journals or working paper series by considering all publication outlets listed in the IDEAS/RePEc ranking. We used the following search terms: "meta-analysis," "meta-regression," "meta-regression analysis," "research synthesis," "systematic review," "quantitative review," "economics," "economics research," "applied economics," and "econometrics." We also used field search terms such as "microeconomics," "macroeconomics," "experimental economics," "industrial relations," "labor economics," and "international economics." The search for data ended July 31, 2021. We included only meta-studies that reported effect sizes with corresponding standard errors and we only included meta-studies that contained at least five primary studies. Where a research area has received more than one meta-analysis or systematic review, we include the most recent and comprehensive study. The list of included meta-studies can be found in the Supporting Information.

The meta-studies can be considered to be representative of the respective research fields that they synthesize but our sample of meta-studies is not necessarily representative of empirical economics. Moreover, meta-studies may analyze primary studies that analyze the same primary data. This is particularly likely to occur in macroeconomics. At the level of meta-analysis, such an overlap could be corrected by using approaches suggested by Bom and Rachinger.[36] In our meta-meta-analysis, we use inference that is clustered at the level of meta-studies to obtain confidence intervals that take the dependence of primary estimates within one meta-study into account. Note that multiple meta-analyses may be published in one meta-study and clustering at the level of meta-studies also takes this dependency into account. This approach follows Abadie et al.[37] and is more conservative than clustering at the level of primary studies and provides wider confidence intervals.

Table 1 provides descriptive information. The number of primary estimates per meta-analysis varies between 4 and 3161 with an average of 367. Meta-analyses in macroeconomics tend to be larger than in microeconomics and the same is true for meta-analysis of observational research as opposed to meta-analysis of experimental research. The sample contains 30 meta-analyses that synthesize exclusively experimental research, such as lab, field, and quasi-experimental designs. A few meta-analyses combine estimates of both experimental and observational research designs and we classified them as observational. We estimate the share of APS as outlined in Ioannidis et al.[3] using 0.8 as the threshold of adequate

TABLE 1 Descriptive statistics.

	N (meta)	*N* (estimates)	Mean	Min	Q25	Q50	Q75	Max
Total	192	70,399	367	4	21	69	480	3161
Microeconomics	131	25,101	192	4	16	42	142	1736
Macroeconomics	61	45,298	743	13	123	525	1092	3161
Experimental	30	1787	60	6	13	19	34	637
Observational	162	68,612	424	4	28	100	600	3161
Share of APS > 0	147	55,622	378	5	24	79	467	3161
Share of APS = 0	45	14,777	328	4	16	35	480	1736

Note: The number of meta-analyses and corresponding primary estimates and the mean, min, max, and quantiles for estimates per meta-analysis are reported. The share of adequately powered studies (APS) is obtained by using a power of 0.8, a significance threshold of 0.05 and weighted least squares to obtain meta-averages.[3]

power and 0.05 as the significance threshold. We split the sample roughly at the median power by considering meta-analysis with at least some APS (Share of APS > 0) and those that have no adequately powered study at all (Share of APS = 0).

2.2 | Counterfactual *p*-values

The observed distribution of *p*-values contains *p*-values that may or may not have been subjected to conscious or unconscious selective reporting. We estimate the distribution of counterfactually expected *p*-values that would have occurred if all studies had estimated the respective genuine effects unbiasedly. Statistical comparison of the observed distribution of published *p*-values with the counterfactual distribution of *p*-values allows us to shed light on the extent of inflated significance in economics. We estimate the distribution of counterfactual *p*-values by using three assumptions that are discussed below. We subject our analysis to sensitivity analyses regarding these three assumptions as outlined in Section 2.3.

Assumption 1. There is one genuine effect per meta-analysis.

Each meta-analysis aims to combine the estimates of primary studies that address the same research question. In economics, most of the variation in published estimates can be usually attributed to methodological heterogeneity rather than heterogeneity in the genuine effect.[13] We conduct sensitivity analysis by allowing for multiple genuine effects per meta-analysis and by using random-effects models and considering the between-study variance in estimating counterfactual *p*-values in Section 2.3.1.

Based on Assumption 1, for all primary estimates holds $E\left[\widehat{\beta}_{ji}\right]=\beta_j$ and thus $\widehat{\beta}_{ji}\sim\mathcal{N}\left(\beta_j,se_{ji}\right)$ where β_j is the true effect in meta-analysis j, $\widehat{\beta}_{ji}$ is the ith estimate of β_j and se_{ji} is the true standard error of estimate i in meta-analysis j. Therefore, the counterfactual z-value (z^{cf}) in the absence of biases would have been drawn from

$$z_{ji}^{cf}\sim\mathcal{N}\left(\frac{\beta_j}{se_{ji}},1\right). \tag{1}$$

The expected frequency of counterfactual z-values in a given interval $[a,b]$ is then given by

$$E\left[r_{a,b}^{cf}\right]=\sum_{j}^{N}\sum_{i}^{M_j}\int_a^b\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(x-\left(\frac{\beta_j}{se_{ji}}\right)\right)^2}{2}\right)dx, \tag{2}$$

where N with $j=1,\ldots,N$ is the number of meta-analyses and M_j with $i=1,\ldots,M_j$ is the number of primary estimates in meta-analysis j. We are interested in counterfactual p-values for two-sided tests as the vast majority of tests reported in economics are two-sided.[38] For two-sided tests, the expected frequency of counterfactual p-values in the interval $[g,h]$ is then given by

$$E\left[f_{g,h}^{cf}\right]=E\left[r_{-b,-a}^{cf}\right]+E\left[r_{a,b}^{cf}\right], \tag{3}$$

where $a=Q(1-h/2)$, $b=Q(1-g/2)$ and Q is the quantile function of the standard normal distribution.

If all studies had estimated the respective genuine effect unbiasedly, $E\left[f_{g,h}^{cf}\right]$ provides the expected frequency of p-values in any p-value interval $[g,h]$. These are counterfactually expected p-values as it is likely that some studies have not estimated the genuine effect unbiasedly. Estimating $E\left[f_{g,h}^{cf}\right]$ requires two further assumptions.

Assumption 2. The genuine effect of each meta-analysis can be approximated by meta-analytical estimators.

For our main analysis, we use a combination of two estimators that are known to reduce bias in the estimation of the underlying genuine effect. Whenever there are at least two adequately powered primary estimates in a given meta-analysis, we use the weighted average of the adequately powered studies (WAAP) which was recently proposed by Ioannidis et al.[3,39] If there are less than two adequately powered estimates in a given meta-analysis, we use the precision-effect test and precision-effect estimate with standard errors (PET-PEESE) procedure that is the most frequently used approach in meta-analyses in economics.[40]

All meta-analytical estimators are known to overestimate the genuine effect if selective reporting is present.[41] Overestimating the genuine effect provides a conservative estimate of the extent of selective reporting, as a larger genuine effect implies a larger probability of obtaining a statistically significant *p*-value for a given research design. Therefore, we report results for the meta-average but also for half the meta-average to show the range implied by this uncertainty. Using half of the meta-average is motivated by recent findings that economic effect sizes are frequently inflated by a factor of two.[3] We further explore sensitivity to alternative meta-analytical estimators in Section 2.3.2 and we use various fractions of the meta-average (Figure S2 in the Supporting Information).

Assumption 3. The standard error of each estimate is not subject to selective reporting itself.

Let us denote the estimated standard error of primary estimate i in meta-analysis j as $\widehat{se}_{ji}$. Researchers that consciously or unconsciously select for statistically significant results are likely to predominantly bias $\widehat{\beta}_{ji}$ rather than $\widehat{se}_{ji}$. However, there are situations in which $\widehat{se}_{ji}$ is biased downwards and this may help to obtain a statistically significant result. For example, a researcher may refrain from using clustered standard errors despite having dependent data. Sensitivity analysis for selective reporting based on biasing $\widehat{se}_{ji}$ is discussed in Section 2.3.3.

We can now estimate the expected frequency of counterfactual p-values that would have occurred if all primary studies had estimated the respective genuine effect unbiasedly, $E\left[f^{cf}_{g,h}\right]$, by plugging in $\widehat{se}_{ji}$ for se_{ji} and $\widehat{\beta}_j$ for β_j in (2). We compare the factual and counterfactual distribution of published p-values by using the relative difference between observed and expected frequencies for a given p-value interval $[g,h]$:

$$D_{g,h} = \frac{f^{f}_{g,h} - E\left[f^{cf}_{g,h}\right]}{f^{f}_{0,1}}, \tag{4}$$

where $f^{f}_{g,h}$ is the factually observed frequency of published p-values in the interval $[g,h]$ and $f^{f}_{0,1}$ is the frequency of observed p-values in the interval $[0,1]$, that is, the total number of published p-values. The difference in relative frequencies allows us to assess the abundance or lack of p-values for any interval along the entire distribution of p-values. For visualization, we rely on z-values that are commonly used in economics for this purpose.[5] Note that p-values and z-values can be transformed into each other and contain the same information.

The main measure of interest is the extent of selective reporting. We define the extent of selective reporting as the difference between observed and expected significant p-values:

$$\text{ESR}_s = f^{f}_{s} - E\left[f^{cf}_{s}\right], \tag{5}$$

where the subscript s denotes statistical significance either at the 0.1 or 0.05 threshold. We will express the extent of selective reporting as the share of significant p-values and as the share of total p-values. A link to the replication package can be found in the Supporting Information.

2.3 | Sensitivity analyses

We explore how our main results depend on the assumptions made in the estimation process. We include alternative assumptions in the sensitivity analyses that may increase or decrease the extent of selective reporting. This transparently provides the reader with the full range of estimates for the extent of selective reporting when considering alternative and plausible assumptions.

2.3.1 | Multiple genuine effects (Assumption 1)

We assume that each meta-analysis comprises primary studies that all estimate the same genuine effect. Variations in the estimated effects between primary studies can be mostly attributed to methodological heterogeneity.[13] Such methodological heterogeneity introduces variation in the estimated effects due to alternative methodological choices, including variations in regression specifications or estimation approaches. It is important to emphasize that we do not want to control for methodological heterogeneity in the estimated effects as it mimics researchers' degrees of freedom in estimating a given genuine effect.[25,26]

However, variation in the estimated effects might be also due to genuine heterogeneity. In this case, it would be false to assume that there is only one genuine effect. Note that funnel plots of meta-analyses is economics predominantly show convergence in the estimated effects as precision increases suggesting one genuine effect to be the common case (for an overview see Stanley and Doucouliagos[13]).

We probe robustness with regard to multiple genuine effects by randomly splitting the primary estimates of each meta-analysis into two groups assuming two genuine effects per meta-analysis. The meta-averages are then calculated using weighted least squares (WLS) for each group. We repeat the random splitting and corresponding estimation of the extent of selective reporting 7500 times. We report the mean, minimum, and maximum extent of selective reporting across these 7500 iterations. We further explore robustness with regard to three genuine effects per meta-analysis by randomly splitting the primary estimates of each meta-analysis into three groups. Additionally, we probe robustness by estimating the genuine effect for each meta-analysis using a random-effects model. We then estimate the counterfactual z-value by considering the between-study variance.[41] Specifically, Equation (1) becomes

$$z_{ji}^{cf} \sim \mathcal{N}\left(\frac{\beta_j}{\sqrt{se_{ji}{}^2 + \tau_j{}^2}}, 1\right), \tag{6}$$

where $\tau_j{}^2$ is the between-study variance of meta-analysis j. Note that our data do not contain information on primary studies. When estimating the random-effects models, we conservatively assume that each primary estimate stems from a separate study. Clustering at the level of meta-studies as described in Section 2.1 accounts for dependency in the data.

2.3.2 | Approximation of the genuine effect (Assumption 2)

Our main analysis uses a combination of WAAP and PET-PEESE. To examine the robustness of the estimate of the extent of selective reporting, we perform additional sensitivity analysis by using PET-PEESE for the entire sample and by using WLS.

We further explore sensitivity by using various fractions of the estimated meta-average to approximate the genuine effect.

2.3.3 | Standard errors (Assumption 3)

Our main analysis takes the estimated standard errors as given and assumes that these estimates were not subject to selection. If, however, these standard errors were subject to selection to obtain statistically significant estimates, the observed standard errors would be biased downwards. We conduct two sensitivity analyses by multiplying the standard errors by 1.5 and 2. If the standard errors are actually larger than observed, then the actual z-values are lower than observed, and the extent of selective reporting is larger.

2.3.4 | Further sensitivity analyses

We also conduct supplemental sensitivity analyses with regard to the estimates that we excluded ($|z| > 20$) by considering $|z| > 50$ and $|z| > 100$ as thresholds of exclusion. We also explore sensitivity by excluding meta-analyses that have more than 90%, 80%, 70%, 60%, and 50% of their primary studies adequately powered, as these large estimates of APS indicate that the meta-average might be overestimated.

3 | RESULTS

We visualize the factual and counterfactual distributions using z-values. These distributions are presented in Figure 1 for the case of the full meta-average. It is striking that both distributions intersect close to the 0.1 threshold of statistical significance ($z = 1.64$) as it is not imposed at any step of our analysis. For the range of non-significant z-values the observed z-values are substantially underrepresented compared to the counterfactual distribution. For the range of significant z-values, the observed z-values are overrepresented. The counterfactual distribution of published z-values looks similar for the case of half the meta-average (see Figure S1 in the Supporting Information).

Visual inspection is supported by Table 2 showing the relative differences in the frequencies of factually observed and counterfactually expected p-values for various intervals of p-values. Throughout the range of non-significant p-values, there is a lack of p-values in each interval that amounts to approximately 2%–5% of the probability mass. There is a tendency that this probability mass decreases toward the threshold of statistical significance with a minimum for the p-value interval of 0.1–0.2 with 2.3% for the full meta-average and 3.1% for half the meta-average. In the range of significant p-values there is an abundance of p-values in each interval that is increasing for smaller p-values. For p-values below 0.001 there is even an abundance of 21.2% of the probability mass for the full meta-average and 29% for half the meta-average. These findings indicate that selective reporting may occur along the entire distribution of p-values. While the lack of p-values in the range of non-significant p-values appears to be rather uniform, an abundance of p-values-appears especially for p-values that are considered to be highly significant.

Our main focus is on the extent of selective reporting and Table 2 presents the extent of selective reporting as a share of all p-values (ESR^{all}) and significant p-values (ESR^{sig}) for the 0.1 and 0.05 levels of statistical significance. For the 0.05 level and the full meta-average, the extent of selective reporting amounts to 32.8% of all p-values and to 57.7% of the p-values that are published as statistically significant. The difference to the 0.1 threshold is small. For half the meta-average and the 0.05 level, the extent of selective reporting increases to 40.8% of all p-values and 71.9% of the significant p-values.

The factual and counterfactual distributions using z-values presented in Figure 2 are broken down by Microeconomics versus Macroeconomics, Experimental versus Observational, and Share of APS > 0 versus Share of APS = 0. This subsample analysis reveals some notable differences that might help to shed light on the underlying determinants of the extent of selective reporting. The factual distributions of published z-values appear similar for Microeconomics and Macroeconomics, while the counterfactual distribution for Macroeconomics tends to have more probability mass in the non-significance

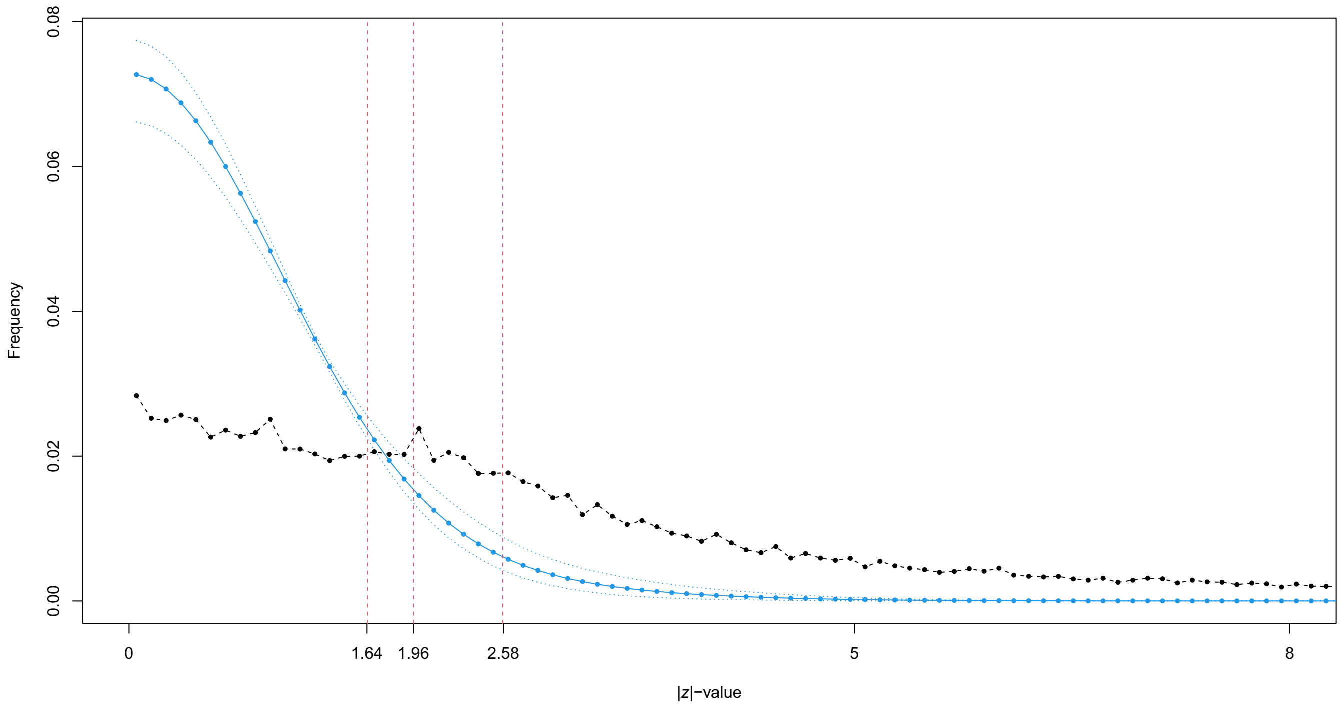

FIGURE 1 Distributions of factual and counterfactual *z*-values. The factual distribution is given by the black dashed line and the counterfactual distribution is given by the blue solid line with dashed 0.95 confidence intervals based on bootstrapping clustered by meta-studies. The dots are placed at the center of each interval using a grid of 0.1025. This interval size ensures that the critical value of 1.64 represents an interval border. The red dashed lines represent the thresholds of statistical significance for the 0.1 ($z = 1.64$), 0.05 ($z = 1.96$) and 0.01 ($z = 2.58$) level.

TABLE 2 Differences in relative frequencies and inflated significance.

	Meta-average		Meta-average/2	
***p*-value interval**	**Difference**	**95% CI**	**Difference**	**95% CI**
$p > 0.9$	−0.039	[−0.048, −0.030]	−0.048	[−0.055, −0.040]
$0.9 > p > 0.8$	−0.041	[−0.050, −0.032]	−0.050	[−0.057, −0.042]
$0.8 > p > 0.7$	−0.042	[−0.050, −0.033]	−0.050	[−0.057, −0.043]
$0.7 > p > 0.6$	−0.041	[−0.049, −0.032]	−0.049	[−0.056, −0.042]
$0.6 > p > 0.5$	−0.043	[−0.051, −0.034]	−0.052	[−0.058, −0.044]
$0.5 > p > 0.4$	−0.038	[−0.046, −0.029]	−0.047	[−0.052, −0.040]
$0.4 > p > 0.3$	−0.034	[−0.042, −0.025]	−0.043	[−0.048, −0.037]
$0.3 > p > 0.2$	−0.034	[−0.041, −0.025]	−0.042	[−0.046, −0.037]
$0.2 > p > 0.1$	−0.023	[−0.030, −0.014]	−0.031	[−0.033, −0.028]
$0.1 > p > 0.05$	0.007	[0.003, 0.012]	0.004	[0.000, 0.007]
$0.05 > p > 0.01$	0.053	[0.045, 0.060]	0.052	[0.040, 0.062]
$0.01 > p > 0.001$	0.063	[0.053, 0.072]	0.067	[0.052, 0.079]
$0.001 > p$	0.212	[0.138, 0.271]	0.290	[0.258, 0.315]
$ESR_{0.10}^{all}$	0.335	[0.256, 0.403]	0.412	[0.353, 0.460]
$ESR_{0.05}^{all}$	0.328	[0.245, 0.397]	0.408	[0.353, 0.454]
$ESR_{0.10}^{sig.}$	0.531	[0.424, 0.623]	0.653	[0.574, 0.720]
$ESR_{0.05}^{sig.}$	0.577	[0.456, 0.684]	0.719	[0.636, 0.792]
No. of meta-analysis	192		192	
No. of tests	70,399		70,399	

Note: 95% confidence intervals based on bootstrapping clustered at the level of meta-studies.

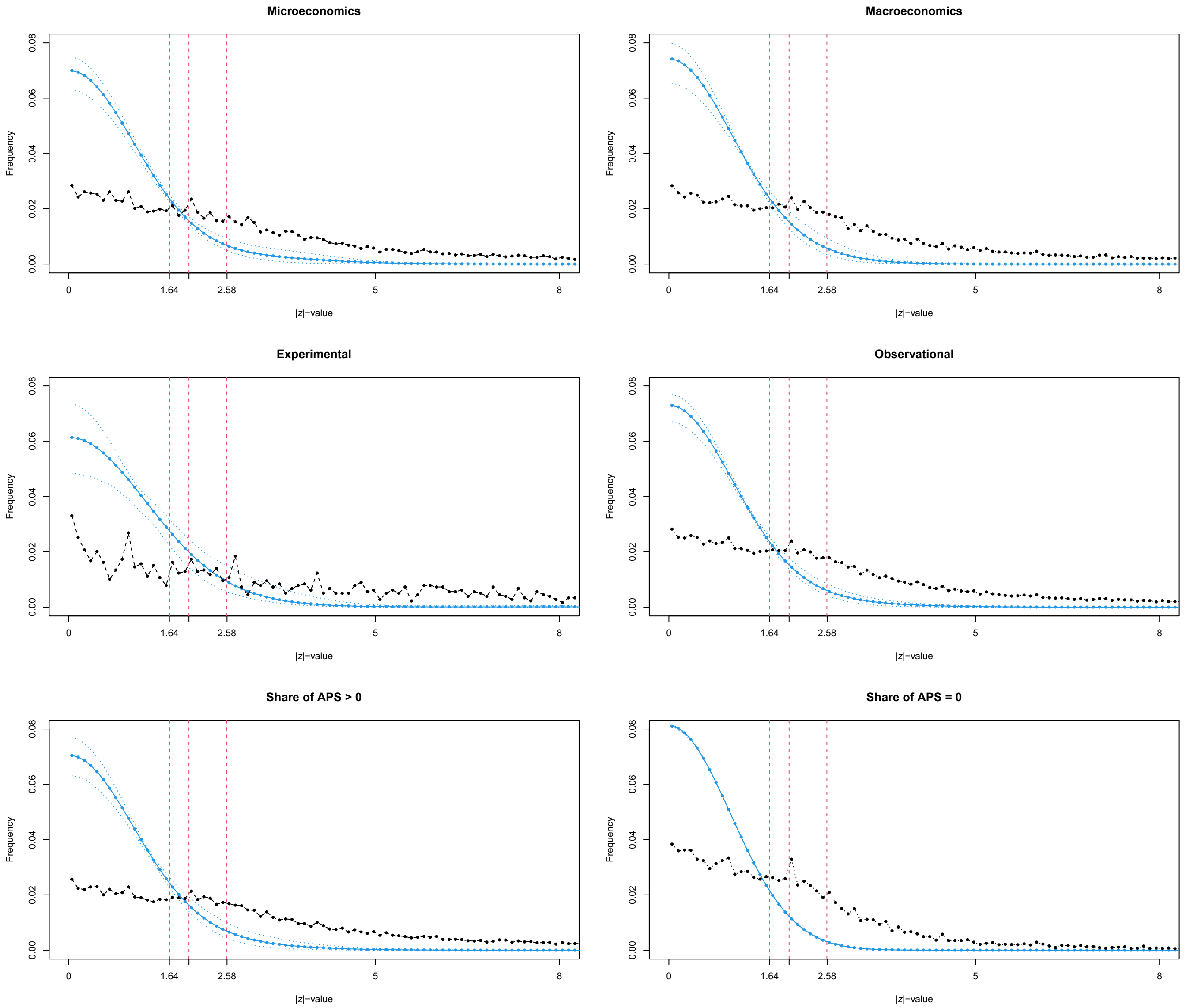

FIGURE 2 Distributions of factual and counterfactual z-values broken down by subsamples. Factual distributions given by the black dashed lines and counterfactual distributions given by the blue solid lines with dashed 0.95 confidence intervals based on bootstrapping clustered by meta-studies. The dots are placed at the center of each interval using a grid of 0.1025. This interval size ensures that the critical value of 1.64 represents an interval border. The red dashed lines represent the thresholds of statistical significance for the 0.1 ($z = 1.64$), 0.05 ($z = 1.96$), and 0.01 ($z = 2.58$) levels.

range. This conveys into an estimated extent of selective reporting as a share of the significant p-values of 50.8% for Microeconomics and 61.5% for Macroeconomics at the 0.05 threshold of significance (see Table S3 in the Supporting Information). One way of interpreting this finding is that more standardized research designs in microeconomics may help to reduce the extent of selective reporting. The confidence intervals are, however, overlapping.

For Experimental versus Observational, the factual and counterfactual distributions of published z-values appear similar for Experimental, which conveys to an extent of selective reporting of 35.7%, but a wider confidence interval (Table S4 in the Supporting Information). Note that the sample contains only 30 meta-analyses of experimental research with 1787 primary estimates. For Observational, the extent of selective reporting is 58.4%.

With regard to Share of APS > 0 versus Share of APS = 0, the factual distributions differ quite substantially. For Share of APS > 0, the factual distribution appears rather flat while for Share of APS = 0 the typical sharp decrease of probability mass in the significance range becomes more apparent. This difference is expected as more APS generate more reliably statistically significant findings. The difference in the extent of selective reporting is striking with 52.6% to 85.6% and the confidence intervals are not overlapping (see Table S5 in the Supporting Information).

Finally, the results for the sensitivity analyses are shown in Figure 3 for the extent of selective reporting as a share of the significant p-values using the 0.05 threshold and the full meta-average. The main result for the full meta-average is reported in A. Differences are small if the meta-average is estimated using exclusively WLS (B) and PET-PEESE (C). The extent of selective reporting increases when random effects models are used and the counterfactual z-values are estimated considering the between-study variance (D). The main finding is also robust when two (E) or three (F) genuine effects are considered. For cases E and F, we report the minimum and maximum of the 7500 iterations rather than the 0.95 bootstrapped confidence intervals. The extent of selective reporting is increased by assuming that the true standard errors are 1.5 times larger (G) or even doubled (H). Doubling the standard errors results in an extent of selective reporting of 72.1%. In all cases, the confidence intervals overlap with the confidence interval of the main estimate (A). Based on these sensitivity analyses the estimated extent of selective reporting appears to be fairly robust with regard to Assumptions 1–3. Additional results for these sensitivity analyses are reported in Tables S6–S9 in the Supporting Information.

Moreover, we explored robustness when z-values larger than 50 (I) and larger than 100 (J) are excluded. Such large z-values are likely to bias the meta-average upwards resulting in smaller estimates of the extent of selective reporting. Even if z-values up to 100 are considered the extent of selective reporting is still estimated to be 43.9%. When we exclude meta-analyses from the sample with more than 90% (K), 80% (L), 70% (M), 60% (N), and 50% (O) of their studies being adequately powered, the extent of selective reporting is 40.7%, 39.4%, 59.1%, 64.6%, and 70%, respectively. The extent of selective reporting seems to be increasing as meta-analyses with either seemingly or truly APS are excluded.

We considered full and half the meta-average for the estimation of the extent of selective reporting in Table 2. The motivation for using half the meta-average is that meta-averages are known to be overestimated if selective

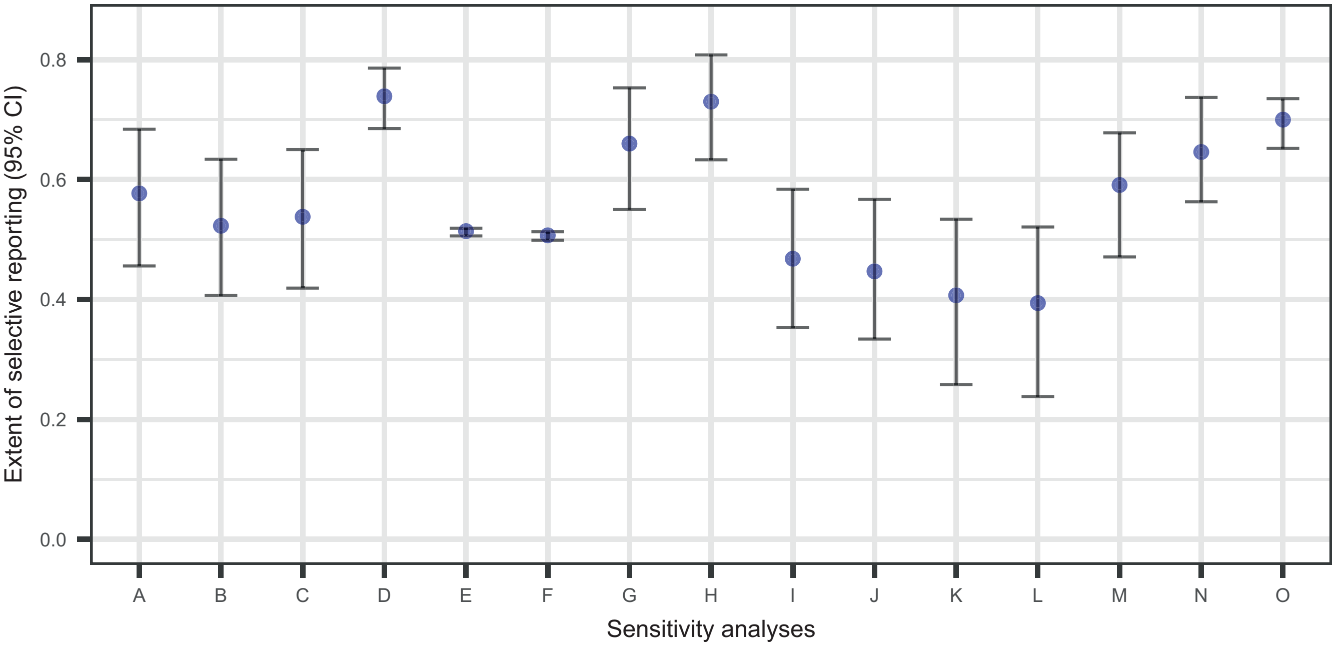

FIGURE 3 Sensitivity analyses of the extent of selective reporting as a share of the significant p-values ($\text{ESR}_{0.05}^{\text{sig}}$). Vertical lines represent 95% confidence intervals. A = Full meta-average, B = WLS, C = PET-PEESE, D = random-effects model and considering between-study variance in estimating counterfactual z-values, E = mean of the extent of selective reporting obtained from 7500 iterations assuming two genuine effects per meta-analysis, F = mean of the extent of selective reporting obtained from 7500 iterations assuming three genuine effects per meta-analysis, G = standard error of each estimate is multiplied by 1.5, H = standard error of each estimate is doubled, I = when $|z| \leq 50$ is considered, J = when $|z| \leq 100$ is considered, K = full meta-average and exclusion of meta-analyses with more than 90% of the primary studies being adequately powered, L = full meta-average and exclusion of meta-analyses with more than 80% of the primary studies being adequately powered, and M = full meta-average and exclusion of meta-analyses with more than 70% of the primary studies being adequately powered, N = full meta-average and exclusion of meta-analyses with more than 60% of the primary studies being adequately powered, and O = full meta-average and exclusion of meta-analyses with more than 50% of the primary studies being adequately powered.

reporting is present. Figure S2 in the Supporting Information provides estimates of the extent of selective reporting for various shares of the full meta-average. As an extreme case, the estimated extent of selective reporting increases to 89.5% if 10% of the full meta-average is considered.

4 | DISCUSSION

By empirically estimating a counterfactual distribution of published p-values that would have occurred if all estimates had estimated the genuine effect unbiasedly, we infer that the extent of selectively reported p-values ranges between 57.7% and 71.9%. This range is supported by comprehensive sensitivity analysis. Specifically, we probe robustness with regard to multiple genuine effects per meta-analysis, alternative approaches to approximate genuine effects, the standard errors being subject to selective reporting, and various alternative data exclusion criteria.

The observed and counterfactually expected distributions of p-values intersect close to the 0.1 threshold of statistical significance, though it is not imposed at any step of our analysis. While the 0.05 threshold is frequently seen as the more important threshold of significance, the 0.1 threshold serves as the first threshold that needs to be passed to emphasize (marginal) statistical significance. This is consistent with the typical eye-catchers used in economics (0.1, 0.05, and 0.01).[2] Moreover, Bruns et al.[5] find evidence for selective reporting at both the 0.1 and 0.05 thresholds for a sample of economics and management articles and the relevance of the 0.1 threshold was also found in other disciplines.[42] Researchers may give "spin" to results that pass the 0.1 threshold to let them appear more significant.[43]

Our analysis quantifies the extent of selectively reported p-values to 57.7% and 71.9% of the significant p-values. Previous research focused on marginally significant p-values and used a theoretical model to infer that 10%–20% of the marginally significant p-values should have not been significant.[44] Our approach allows us to infer the abundance and lack of p-values along the entire distribution of p-values. Our findings suggest that p-values are rather uniformly missing in the non-significance range and predominantly shifted to highly significant p-values ($p < 0.001$). Bruns and Ioannidis[15] illustrate how selective reporting in observational research can result in very small p-values. If indeed bias drives the p-values to very low levels, such as $p < 0.001$, then the proposed solution of shifting the threshold of statistical significance to $p < 0.005$ instead of $p < 0.05$[45] would not suffice to address spurious significance in economics. Measures such as pre-registration and availability of raw data may help understand and curtail some of these biases.

Our subsample analysis suggests that experimental research designs tend to exhibit less selective reporting. A key factor that influences the extent of selective reporting is the share of APS in a given literature. This finding corroborates the analysis by Ioannidis et al.[3] who find that economics research is severely underpowered. Moreover, microeconomics tends to be less prone to selective reporting than macroeconomics which might be related to more standardized research designs, that is, research designs with less researchers' degrees of freedom.

We also perform tests for the presence of selective reporting which corroborates our findings. These tests are designed to diagnose the presence of selective reporting but not to estimate its extent. First, we conduct the Caliper test as proposed by Gerber and Malhotra.[46] This test explores whether marginally significant z-values are overrepresented compared to marginally non-significant z-values. The Caliper test indicates the presence of selective reporting at the 0.05 threshold of significance and to some extent at the 0.01 threshold (Table S1 in the Supporting Information). Second, we also apply the recently proposed tests by Elliott et al.[47] and these tests also indicate the presence of selective reporting (see Table S2 in the Supporting Information).

We applied the counterfactual p-value approach to economics. This approach is based on the analysis of a large set of meta-analyses and opens new avenues to study patterns and mechanisms of selective reporting across disciplines by analyzing the abundance and lack of p-values throughout the entire p-value distribution. Meta-meta-analyses become increasingly popular in the assessment of selective reporting. For example, Yang et al.[48] use 87 meta-analyses to show that findings in ecology and evolutionary biology are inflated, Bartos et al.[49] apply Robust Bayesian Meta-Analysis (RoBMA) to 90 meta-analyses to show that psychological meta-analyses frequently overestimate the presence of a meta-analytic effect and its magnitude, and Fanelli et al.[9] use 3042 meta-analyses to assess bias patterns and risk factors across disciplines.

While we estimate the extent of selective reporting to be large, it should be emphasized that we cannot conclude anything about the existence or absence of genuine effects in the respective research fields. Researchers may inflate statistical significance by exaggerating effect sizes[3] but that does not mean that a genuine effect is absent. It is likely that the analyzed effects are (much) smaller than what is suggested in the literature. Moreover, as already emphasized by Ioannidis et al.[43] we cannot distinguish between different types of selective reporting (p-hacking, HARKing, and publication bias).

The analysis in this article was conducted at the level of p-values or z-values, respectively. Alternatively, the analysis could be also conducted at the level of primary studies by weighting each p-value with the inverse number of p-values in the respective article. The abundance of highly significant p-values might become smaller when analyzing at the level of primary studies, if primary studies with small p-values systematically contribute more estimates compared to studies with larger p-values. Primary studies with large estimates and correspondingly small p-values might easily report multiple estimates as statistical significance is easily achieved while primary studies with small estimates might struggle to provide many statistically significant estimates. This is an interesting avenue for future research.

Different meta-analyses may differ on the extent to which they may have incorporated also some gray literature through systematic searches of gray literature sources. Roughly 20% of the considered estimates stem from gray literature and authors normally provide a rationale to why gray literature is not considered for a specific research question. In theory, considering gray literature may affect the extent of selective reporting reflected in each meta-analysis. However, in the typical meta-analysis, in the absence of study pre-registration it is impossible to know for sure how many studies are not published.

Another issue is that meta-analyses may sometimes focus on effects that were not those of primary interest in some studies. It is possible that selective reporting bias may affect less effects that are not of primary interest. However, this is uncertain and, in fact, in some scientific fields (e.g., medicine) it has been shown that "gold standard" large trials tend to agree more with meta-analyses of primary outcomes than with meta-analyses of secondary outcomes, suggesting that selection biases may be even larger for secondary than for primary outcomes.[50]

5 | CONCLUSION

Using a large sample of 70,399 published p-values from 192 meta-analyses, we infer that 57.7%–71.9% less p-values should have been published as being statistically significant. This range is supported by comprehensive sensitivity analysis. Subsample analyses indicate that the extent of selective reporting could be related to research designs and statistical power. While economics research has already shifted attention to experimental[33] and quasi-experimental research,[51] explicit power considerations in both (quasi-)experimental and observational research receive less attention. Corroborating Ioannidis et al.,[3] we advocate emphasis on routinely considering statistical power to avoid the temptation to exaggerate effect sizes with the aim of statistical significance.

AUTHOR CONTRIBUTIONS

Stephan B. Bruns: Conceptualization; methodology; writing – original draft; writing – review and editing; formal analysis. **Teshome K. Deressa:** Methodology; formal analysis; writing – review and editing. **T. D. Stanley:** Methodology; writing – review and editing. **Chris Doucouliagos:** Data curation; writing – review and editing. **John P. A. Ioannidis:** Methodology; writing – review and editing.

ACKNOWLEDGMENTS

We thank seminar participants at the Meta-Research Innovation Center at Stanford, Hasselt University, University of Göttingen, and the MAER-Net Colloquium 2017 for helpful comments on an earlier version. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the supplementary material of this article.

ORCID

Stephan B. Bruns https://orcid.org/0000-0002-3028-9699
T. D. Stanley https://orcid.org/0000-0002-3205-1983

ENDNOTES

[1] We excluded 4141 observations with absolute z-values being larger than 20 to reduce the risk that estimated meta-averages are influenced by large outliers. Sensitivity analysis also explores thresholds of 50 and 100.

[2] For example, Puetz and Bruns[38] collect eye-catchers for a large sample of economics articles and find that the typical thresholds used in economics tables are the 0.1, 0.05, and 0.01 thresholds.

REFERENCES

1. Havránek T. Measuring intertemporal substitution: the importance of method choices and selective reporting. *J Eur Econ Assoc*. 2015;13(6):1180-1204.
2. Camerer CF, Dreber A, Forsell E, et al. Evaluating replicability of laboratory experiments in economics. *Science*. 2016; 351(6280):1433-1436. doi:10.1126/science.aaf0918
3. Ioannidis J, Stanley T, Doucouliagos C. The power of bias in economics research. *Econ J*. 2017;127:F236-F265.
4. Chang AC, Li P. A preanalysis plan to replicate sixty economics research papers that worked half of the time. *Am Econ Rev*. 2017;107(5):60-64.

5. Bruns SB, Asanov I, Bode R, et al. Reporting errors and biases in published empirical findings: evidence from innovation research. *Res Policy*. 2019;48(9):103796.
6. Vivalt E. Specification searching and significance inflation across time, methods and disciplines. *Oxf Bull Econ Stat*. 2019; 81(4):797-816.
7. Havranek T, Irsova Z, Laslopova L, Zeynalova O. Publication and attenuation biases in measuring skill substitution. *Rev Econ Stat*. 2022;1-37. doi:10.1162/rest_a_01227
8. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015;349(6251):253-267. doi:10.1126/science.aac4716
9. Fanelli D, Costas R, Ioannidis JPA. Meta-assessment of bias in science. *Proc Natl Acad Sci*. 2017;114(14):3714-3719. doi:10.1073/pnas.1618569114
10. David SP, Naudet F, Laude J, et al. Potential reporting bias in neuroimaging studies of sex differences. *Sci Rep*. 2018;8(1):1-8.
11. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
12. Glaeser EL. Researcher incentives and empirical methods. In: Caplin A, Schotter A, eds. *The Foundations of Positive and Normative Economics: A Handbook*. Oxford University Press; 2008: 300-319.
13. Stanley TD, Doucouliagos H. *Meta-Regression Analysis in Economics and Business*. Routledge; 2012.
14. Simonsohn U, Nelson LD, Simmons JP. P-curve: a key to the file-drawer. *J Exp Psychol Gen*. 2014;143(2):534-547. doi:10.1037/a0033242
15. Bruns SB, Ioannidis JPA. P-curve and p-hacking in observational research. *PLoS One*. 2016;11(2):e0149144. doi:10.1371/journal.pone.0149144
16. Kerr NL. HARKing: hypothesizing after the results are known. *Pers Soc Psychol Rev*. 1998;2(3):196-217. doi:10.1207/s15327957pspr02034
17. Rosenthal R. The file drawer problem and tolerance for null results. *Psychol Bull*. 1979;86(3):638-641. doi:10.1037/0033-2909.86.3.638
18. Nosek BA, Spies JR, Motyl M. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect Psychol Sci*. 2012;7(6):615-631.
19. Leamer EE, Leamer EE. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Vol 53. Wiley; 1978.
20. Hendry DF. Econometrics – alchemy or science? *Economica*. 1980;47(188):387-406. doi:10.1093/0198293542.001.0001
21. Sterling TD. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J Am Stat Assoc*. 1959;54(285):30-34.
22. Selvin HC, Stuart A. Data-dredging procedures in survey analysis. *Am Stat*. 1966;20(3):20-23.
23. Leamer EE. Sensitivity analysis would help. *Am Econ Rev*. 1985;75(3):308.
24. Bruns SB, Kalthaus M. Flexibility in the selection of patent counts: implications for p-hacking and evidence-based policymaking. *Res Policy*. 2020;49(1):103877.
25. Huntington-Klein N, Arenas A, Beam E, et al. The influence of hidden researcher decisions in applied microeconomics. *Econ Inq*. 2021;59(3):944-960.
26. Silberzahn R, Uhlmann EL, Martin DP, et al. Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv Methods Pract Psychol Sci*. 2018;1(3): 337-356.
27. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci*. 2011; 22(11):1359-1366. doi:10.1177/0956797611417632
28. Casey K, Glennerster R, Miguel E. Reshaping institutions: evidence on aid impacts using a preanalysis plan. *Q J Econ*. 2012; 127(4):1755-1812. doi:10.1093/qje/qje027
29. Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: unlocking the file drawer. *Science*. 2014; 345(6203):1502-1505. doi:10.1126/science.1255484
30. Kunda Z. The case for motivated reasoning. *Psychol Bull*. 1990; 108(3):480-498. doi:10.1037/0033-2909.108.3.480
31. Gechert S, Havranek T, Irsova Z, Kolcunova D. Measuring capital-labor substitution: the importance of method choices and publication bias. *Rev Econ Dyn*. 2022;45:55-82.
32. Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clin Trials*. 2007;4(3):245-253.
33. Duflo E, Banerjee AV, Finkelstein A, Katz LF, Olken B, Sautmann A. In praise of moderation: suggestions for the scope and use of pre-analysis plans for RCTs in economics. NBER Working Paper. 2020 (w26993).
34. Olken BA. Promises and perils of pre-analysis plans. *J Econ Perspect*. 2015;29(3):61-80.
35. Imbens GW. Statistical significance, p-values, and the reporting of uncertainty. *J Econ Perspect*. 2021;35(3):157-174.
36. Bom PR, Rachinger H. A generalized-weights solution to sample overlap in meta-analysis. *Res Synth Methods*. 2020;11(6):812-832.
37. Abadie A, Athey S, Imbens GW, Wooldridge J. When should you adjust standard errors for clustering? *Q J Econ*. 2023;138:1-35.
38. Pütz P, Bruns SB. The (non-)significance of reporting errors in empirical economics: evidence from three top journals. *J Econ Surveys*. 2021;35(1):348-373.
39. Kraemer HC, Gardner C, Brooks JO III, Yesavage JA. Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychol Methods*. 1998;3(1):23-31.
40. Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. *Res Synth Methods*. 2014; 5(1):60-78.
41. Stanley T, Doucouliagos H, Ioannidis JP, Carter EC. Detecting publication selection bias through excess statistical significance. *Res Synth Methods*. 2021;12:1-20.
42. Kavvoura FK, McQueen MB, Khoury MJ, Tanzi RE, Bertram L, Ioannidis JP. Evaluation of the potential excess of statistically significant findings in published genetic association studies: application to Alzheimer's disease. *Am J Epidemiol*. 2008;168(8):855-865.
43. Ioannidis JP. Clarifications on the application and interpretation of the test for excess significance and its extensions. *J Math Psychol*. 2013;57(5):184-187.
44. Brodeur A, Lé M, Sangnier M, Zylberberg Y. Star wars: the empirics strike back. *Am Econ J Appl Econ*. 2016;8(1):1-32. doi:10.1257/app.20150044
45. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav*. 2018;2(1):6-10.
46. Gerber S, Malhotra N. Publication bias in empirical sociological research: do arbitrary significance levels distort published results? *Sociol Methods Res*. 2008;37(1):3-30. doi:10.1177/0049124108318973
47. Elliott G, Kudrin N, Wüthrich K. Detecting p-hacking. *Econometrica*. 2022;90(2):887-906.

48. Yang Y, Sánchez-Tójar A, O'Dea RE, et al. Publication bias impacts on effect size, statistical power, and magnitude (Type M) and sign (Type S) errors in ecology and evolutionary biology. *BMC Biol*. 2023;21(1):1-20.
49. Bartoš F, Maier M, Shanks DR, Stanley T, Sladekova M, Wagenmakers EJ. Meta-analyses in psychology often overestimate evidence for and size of effects. *R Soc Open Sci*. 2023;10(7):230224.
50. Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. *JAMA*. 1998;279(14):1089-1093.
51. Angrist JD, Pischke JS. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *J Econ Perspect*. 2010;24(2):3-30.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bruns SB, Deressa TK, Stanley TD, Doucouliagos C, Ioannidis JPA. Estimating the extent of selective reporting: An application to economics. *Res Syn Meth*. 2024;1-13. doi:10.1002/jrsm.1711