












# Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics

František Bartoš<sup>1,2</sup>  | Maximilian Maier<sup>3</sup>  | Eric-Jan Wagenmakers<sup>1</sup>  |  
 Franziska Nippold<sup>4</sup> | Hristos Doucouliagos<sup>5</sup>  | John P. A. Ioannidis<sup>6,7,8,9,10</sup>  |  
 Willem M. Otte<sup>11</sup>  | Martina Sladekova<sup>12</sup>  | Teshome K. Deressa<sup>13</sup>  |  
 Stephan B. Bruns<sup>6,13,14</sup>  | Daniele Fanelli<sup>15,16</sup>  | T. D. Stanley<sup>5</sup> 

<sup>1</sup>Department of Psychological Methods, University of Amsterdam, Amsterdam, Netherlands

<sup>2</sup>Institute of Computer Science, Czech Academy of Sciences, Prague, Czech Republic

<sup>3</sup>Department of Experimental Psychology, University College London, London, UK

<sup>4</sup>Department of Psychology, University of Amsterdam, Amsterdam, Netherlands

<sup>5</sup>Department of Economics, Deakin University, Geelong, Victoria, Australia

<sup>6</sup>Meta-Research Innovation Center at Stanford (METRICS), Stanford, California, USA

<sup>7</sup>Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, California, USA

<sup>8</sup>Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, California, USA

<sup>9</sup>Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, California, USA

<sup>10</sup>Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, California, USA

<sup>11</sup>Department of Pediatric Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

<sup>12</sup>School of Psychology, University of Sussex, Sussex, UK

<sup>13</sup>Centre for Environmental Sciences, Hasselt University, Hasselt, Belgium

<sup>14</sup>Department of Economics, University of Göttingen, Göttingen, Germany

<sup>15</sup>Department of Methodology, London School of Economics and Political Science, London, UK

<sup>16</sup>Doctoral Centre, School of Social Sciences, Heriot-Watt University, Edinburgh, UK

## Correspondence

František Bartoš, Department of Psychological Methods, University of Amsterdam, Amsterdam, Netherlands.  
 Email: [f.bartos96@gmail.com](mailto:f.bartos96@gmail.com)

## Funding information

Nederlandse Organisatie voor Wetenschappelijk Onderzoek; Special Research Fund (BOF) of Hasselt University, Grant/Award Number: BOF20OWB05; German Research

## Abstract

Publication selection bias undermines the systematic accumulation of evidence. To assess the extent of this problem, we survey over 68,000 meta-analyses containing over 700,000 effect size estimates from medicine (67,386/597,699), environmental sciences (199/12,707), psychology (605/23,563), and economics (327/91,421). Our results indicate that meta-analyses in economics are the most severely contaminated by publication selection bias, closely followed by meta-analyses in environmental sciences and psychology, whereas meta-analyses in medicine are contaminated the least. After adjusting for publication selection bias, the median probability of the presence

František Bartoš and Maximilian Maier have contributed equally.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Research Synthesis Methods* published by John Wiley & Sons Ltd.

Foundation (DFG), Grant/Award  
Number: 405039391

of an effect decreased from 99.9% to 29.7% in economics, from 98.9% to 55.7% in psychology, from 99.8% to 70.7% in environmental sciences, and from 38.0% to 29.7% in medicine. The median absolute effect sizes (in terms of standardized mean differences) decreased from  $d = 0.20$  to  $d = 0.07$  in economics, from  $d = 0.37$  to  $d = 0.26$  in psychology, from  $d = 0.62$  to  $d = 0.43$  in environmental sciences, and from  $d = 0.24$  to  $d = 0.13$  in medicine.

#### KEYWORDS

Bayesian, effect sizes, evidence, meta-analysis, model-averaging, publication bias, RoBMA

#### Highlights

##### What is already known

- Publication selection bias, where studies with significant or positive results are more likely to be reported and published, distorts the available scientific record.

##### What is new

- This study surveyed over 68,000 meta-analyses from medicine, environmental sciences, psychology, and economics to assess the extent of publication selection bias. As a result, it underscores the importance of addressing publication bias in evidence synthesis.
- Results suggest that meta-analyses in economics are the most affected by publication selection bias, followed by environmental sciences and psychology. In contrast, meta-analyses in medicine are suggested to be the least affected. Yet, notable biases are found across all of these scientific disciplines.

##### Potential impact for readers

- This study documents the potential extent of publication bias in different fields, which could help researchers and the public better understand the limitations of research and the potential biases of research synthesis.

## 1 | INTRODUCTION

Publication selection biases (PSB) are defined as the selective reporting of results in ways that deviate from the objective, complete scientific record. PSB may entail the suppression of “negative” findings or the conversion of “negative” results into more “positive” ones (e.g., those with more favorable  $p$ -values and/or with larger effect sizes) and might represent a problem in all scientific disciplines, for example, References 1–5. Studies that examine the self-reported behavior of researchers show that 78% of researchers failed to report all dependent measures of a study<sup>6</sup> (however, see,<sup>7</sup> for a response that suggests a lower proportion). Some studies also suggest that PSB might be modestly increasing in some areas, although the exact nature, prevalence, and impact of PSB is unknown and likely to be variable across scientific fields.<sup>8,9</sup>

To gauge the extent of the PSB, one would need to have access to the complete scientific record or a

representative and wide-coverage sample of it. However, this is infeasible as much of the relevant data is not publicly recorded. Instead, the footprint of PSB is indirectly probed by re-analyzing meta-analyses in several specific fields with different statistical techniques<sup>9–14</sup> and focusing on patterns in the published results that would herald the presence of PSB. All these available methods try to identify the footprint of PSB, and thus their results need to be interpreted with caution since these patterns (e.g., correlations of effect sizes and standard errors) may sometimes be due to factors other than PSB (e.g., genuine heterogeneity across studies). However, when large numbers of meta-analyses show the same patterns, this constitutes a probable footprint of PSB, which can be used to estimate its relative magnitude across different fields.

Previous field-wide assessments of PSB suggested that the prevalence of over-reporting positive results and other possible symptoms of bias increased moving from

the physical to the biological and the social sciences, and even suggested that problems might be worsening over time in the latter.<sup>9,15–20</sup> However, these estimates were based on proxy measures of PSB that have several limitations.

To our knowledge, no previous survey of the potential footprint of PSB has used state-of-the-art methods. Our proposed approach is more comprehensive than past surveys: employing different strategies to identify potential PSB, using new measures of PSB, and analyzing a much larger number of research studies covering the fields of medicine, environmental sciences, psychology, and economics.

## 2 | METHODS

### 2.1 | Data sets

We used five large data sets from medicine, environmental sciences, psychology, and economics. The data set from medicine comprises meta-analyses of continuous and dichotomous outcomes obtained from the Cochrane Database of Systematic Reviews published between 1997 and 2020. The data set from environmental sciences comprises the meta-analyses of mean differences, odds ratios, and correlation coefficients by Deressa et al.<sup>21</sup> published between 2010 and 2020. The data sets from psychology comprise the meta-analyses of mean differences and correlation coefficients by Stanley et al.<sup>12</sup> published between 2011 and 2016 combined with a random sample of meta-analyses published in psychological journals by Sladekova et al.<sup>22</sup> published in 2008 and 2018. Finally, the data set from economics comprises the extended data set of meta-analyses of regression and correlation coefficients by Ioannidis et al.<sup>10</sup> published between 1967 and 2021. Eighty-four meta-analyses were part of both the Ioannidis et al.<sup>10</sup> and Stanley et al.<sup>12</sup> data set. Since each of the meta-analyses could be classified in both fields (psychology or economics), we did not remove them from either of the data sets. From each data set we only used meta-analyses with at least three estimates reported using standardized effect size metrics such as log odds ratios, standardized mean differences, and (partial) correlation coefficients that can be transformed to a common standardized mean difference effect size metric, Cohen's *d*.

#### 2.1.1 | Medicine

The data set from medicine comprises meta-analyses of continuous and dichotomous outcomes obtained from

the Cochrane Database of Systematic Reviews (CSDR) published between 1997 and 2020. We identified systematic reviews in the CDSR through PubMed, limiting the period to Jan 2000–May 2020. For that, we used the NCBI's EUtils API with the following query: “Cochrane Database Syst Rev”[journal] AND (“2000/01/01” [PDAT]: “2020/05/31” [PDAT]). For each review, we downloaded the XML meta-analysis table file (rm5-format) associated with the review's latest version. We extracted the tables with continuous and dichotomous outcomes from these rm5 files with a custom Javascript and R programs (<https://github.com/wmotte/cochrane2022>).

We selected meta-analysis tables based on the highest aggregation reported in the CSDR. For each meta-analysis, we removed estimates based on one or fewer participants in the control or treatment group and used all meta-analyses with at least three effect size estimates.

#### 2.1.2 | Environmental sciences

The environmental sciences data set consists of meta-analyses of mean differences, correlation coefficients, and odds ratios published between 2010 and 2020. The literature search was performed in the Scopus database using the query: “TITLE-ABS-KEY (“meta analy\*” OR “meta-analy\*” OR “metaanaly\*” OR “meta reg\*” OR “meta-reg\*” OR “metareg\*”) AND SUBJAREA (envi)” on July 21, 2020. Detailed information about the sampling strategy and inclusion/exclusion criteria used can be found in Deressa et al.<sup>21</sup>

#### 2.1.3 | Psychology

The data set from psychology comprises the data set of meta-analyses of mean differences and correlation coefficients of Stanley et al.<sup>12</sup> published between 2011 and 2016 combined with data from Sladekova et al.,<sup>22</sup> a random sample of 433 meta-analyses from 90 articles published in 2008 and 2018. See Stanley et al.<sup>12</sup> and Sladekova et al.<sup>22</sup> for more details about the collected data sets. None of the meta-analyses by Sladekova et al.<sup>22</sup> were published in Psychological Bulletin, precluding overlap with Stanley et al.<sup>12</sup> data set.

#### 2.1.4 | Economics

The data set from economics comprises the extended data set of meta-analyses of regression and correlation coefficients of Ioannidis et al.<sup>10</sup> published between 1967 and 2021. The meta-analyses were identified using various

search engines (e.g., Econlit and Scopus), publisher sites (e.g., Science Direct, Sage, and Wiley), webpages of researchers known to publish meta-analyses, and by searching all volumes of individual journals that are known to publish meta-analyses. We also emailed 109 research teams (associated with either sole-authored or co-authored meta-analyses) for data, with a 67% response rate. The search for data ended on May 30th, 2021.

We selected meta-analyses of standardized mean differences, (partial) correlation coefficients, and mean differences (if enough information was available to compute the standardized mean differences).

### 2.1.5 | Effect size calculation

In cases where the data set did not already feature standardized effect size (Cohen's  $d$ , correlation coefficient  $r$ ,  $\log(OR)$ , or Fisher's  $z$ ), we used the `metafor` R package<sup>23</sup> to calculate the standardized effect sizes. For dichotomous outcomes with zero cell counts, we used the default empty cell correction, adding 1/2 to empty cells. Finally, we converted all standardized effect sizes to Fisher's  $z$  by using the formulas in Borenstein et al.<sup>24</sup>

## 2.2 | Publication bias adjustment with Bayesian model-averaging

We used the PSB detection and correction technique RoBMA-PSMA.<sup>25,26</sup> RoBMA employs Bayesian model-averaging<sup>27,28</sup> and combines the best of two well-performing publication bias adjustment methods: selection models with six different weight functions that adjust for publication selection across a combination of statistical significance and direction of the effect<sup>29</sup> and PET-PEESE, which adjusts for the relationship between effect sizes and standard errors or standard errors squared.<sup>30</sup> Bayesian model-averaging allows us to combine these publication bias adjustment methods based on their predictive adequacy, such that models that predict well have a larger impact on the inference. In that way, we can evaluate the evidence in favor or against the hypothesis of PSB and its impact without committing to any single estimation or correction method.<sup>27</sup>

We used the default RoBMA parameterization which was shown to achieve better performance in both simulation studies and real data examples than either of publication bias adjustment methods alone.<sup>26</sup> It gives equal prior model probabilities to models assuming the presence vs. absence of an effect, heterogeneity, and publication selection bias. RoBMA employs a standard normal

distribution on the effect size,  $\mu \sim \text{Normal}(0, 1)$ , empirically informed Inverse-gamma distribution on the heterogeneity,  $\tau \sim \text{Inverse-Gamma}(1, 0.15)$ ,<sup>31</sup> cumulative unit Dirichlet prior distributions on publication probabilities, and Cauchy prior distributions on the PET-PEESE regression coefficients,  $\text{PET} \sim \text{Cauchy}_+(0, 1)$ ,  $\text{PEESE} \sim \text{Cauchy}_+(0, 5)$ .

### InfoBox 1: Bayes factors

The Bayes factor is the key inference criterion for much of Bayesian statistics, for example Reference 32,33. It compares the relative predictive accuracy (i.e., likelihood of the data) under competing hypotheses (e.g.,  $\mathcal{H}_1$  vs.  $\mathcal{H}_0$ ) and it can also be expressed as the ratio of prior and posterior model odds,

$$\underbrace{\frac{p(\text{data}|\mathcal{H}_1)}{p(\text{data}|\mathcal{H}_0)}}_{\text{Bayes factor}} = \underbrace{\frac{p(\mathcal{H}_1|\text{data})}{p(\mathcal{H}_0|\text{data})}}_{\text{Posterior odds}} \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\text{Prior odds}}$$

Although the Bayes factor is a continuous measure of strength of evidence, the following rules of thumb may aid interpretation: Bayes factors between 1 and 3 are commonly regarded as weak evidence, Bayes factors between 3 and 10 as moderate evidence, and Bayes factors larger 10 as strong evidence for the alternative (or the hypothesis at the top of Equation (1)). When the evidence for the null is considered, the Bayes factor is simply inverted. In other words, a Bayes factor between 1/3 and 1 is considered weak evidence, a Bayes factor between 1/10 and 1/3 moderate and smaller 1/10 strong evidence for the null (e.g., References 34, 35).

## 2.3 | Measures

For each meta-analysis, we used RoBMA to calculate the (PSB) adjusted posterior model-averaged effect size assuming it is present (i.e., without averaging over the point null models to reduce shrinkage toward zero),  $\mu_{\text{adj},k}$ ; publication bias adjusted posterior probability of the presence of the effect,  $p_{\text{adj},k}(\mathcal{H}_1|\text{data}_k)$ ; and the posterior probability of the presence of PSB,  $p_{\text{adj},k}(\mathcal{H}_{\text{psb}}|\text{data}_k)$ . To isolate the effect of PSB adjustment, we compare the Bayesian, PSB unadjusted, model-averaged meta-analysis by dropping the PSB adjustment and thereby estimating the unadjusted posterior probability of the presence of the effect assuming it is present,  $p_{\text{unadj},k}(\mathcal{H}_1|\text{data}_k)$ .

$k = 1, \dots, K$  to denotes the individual meta-analyses. Each meta-analysis is based on  $N_k$  estimates that are characterized with data describing the effect size  $y_{k,n}$  and standard error  $se_{k,n}$ .

### 2.3.1 | Evidence for the effect

We used the change in the posterior probability of the effect and the (standardized) evidence inflation factor to quantify the effect of PSB on meta-analytic evidence.

The posterior probability of the effect is an intuitive way of quantifying the evidence in favor of the alternative hypothesis of the presence of an effect. Under the assumption of equal prior probability of the presence and the absence of the effect,  $p(\mathcal{H}_1) = p(\mathcal{H}_0)$ , posterior probabilities larger than 0.5 indicate that the data are more likely under the presence of the effect. On the other hand, posterior probabilities lower than 0.5 indicate that the data are more likely under the absence of the effect. The ability to quantify evidence for both the null and the alternative is a key benefit of Bayesian methods over null hypothesis significance testing.<sup>36,37</sup>

A corresponding way of quantifying the evidence of an effect is via Bayes factors (see InfoBox 1 for more detail). Bayes factors quantify the change from prior to posterior odds for the presence of the effect. The advantage of Bayes factors is that they are independent of the prior odds for the presence of the effect. In other words, Bayes factors isolate the evidence for the presence of the effect contained in the data. In our settings, the assumption of equal prior probabilities leads to an equivalence between Bayes factors and posterior odds.

The change from the PSB unadjusted posterior probability of the effect,  $p_{\text{unadj},k}(\mathcal{H}_1 | \text{data}_k)$ , to the PSB adjusted posterior probability of the effect,  $p_{\text{adj},k}(\mathcal{H}_1 | \text{data}_k)$ , quantifies the amount of evidence introduced by PSB. The larger the impact of PSB, the larger the difference between the PSB unadjusted and PSB adjusted posterior probabilities of the effect. If there was no PSB, we would observe no change in the posterior probability of the effect after PSB adjustment.

Evidence inflation factor (EIF) quantifies the degree to which the evidence in favor of the presence of the effect was inflated due to PSB.  $\text{EIF}_k$  quantifies the amount of evidence in favor of the effect in the PSB unadjusted meta-analysis,  $\text{BF}_{10,\text{unadj},k}$ , to the amount of evidence in favor of the effect in the PSB adjusted meta-analysis  $\text{BF}_{10,\text{adj},k}$ ,

$$\text{EIF}_k = \frac{\text{BF}_{10,\text{unadj},k}}{\text{BF}_{10,\text{adj},k}}. \quad (1)$$

An evidence inflation factor larger than one indicates inflated evidence in favor of the effect due to PSB.

However, the amount of evidence contained in each meta-analysis, and the corresponding evidence inflation, is dependent on the number of meta-analyzed estimates,  $N_n$ , that is, more estimates lead to more evidence. To facilitate the comparison of evidence inflation due to PSB in meta-analyses with different numbers of estimates, we also compute the standardized, per-estimate, evidence inflation factor in each meta-analysis,  $\text{sEIF}_k$ , by standardizing the EIF by the number of estimates,

$$\text{sEIF}_k = \text{EIF}_k^{\frac{1}{N_k}}, \quad (2)$$

where sEIF represents each estimate's marginal contribution, on average, to the evidence inflation due to PSB. The sEIF also partially mitigates the potential issue of dependent estimates within a meta-analysis. In the most extreme case, e.g., identical estimates, the same data is conditioned upon multiple times, which leads to overestimation of evidence. Taking only a fraction of each estimate's likelihood, proportional to the number of estimates, then ensures that the data are not conditioned upon more than once, although data producing multiple estimates are still weighted more heavily.

### 2.3.2 | Effect size estimates

Absolute bias (bias) quantifies the degree to which the average effect sizes in each meta-analysis,

$$\hat{y}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} y_{k,n},$$

overestimates the PSB-adjusted meta-analytic effect size estimate assuming the presence of the effect  $\mu_{\text{adj},k}$ ,

$$\text{bias}_k = \hat{y}_k - \mu_{\text{adj},k}. \quad (3)$$

Absolute bias larger than zero indicates that PSB leads to inflated effect size estimates. We compare the average effect sizes to the PSB-adjusted effect sizes assuming the presence of the effect (conditional effect size estimates) rather than averaging across all models. Excluding models assuming the absence of a mean effect mitigates the pooling toward 0 in meta-analyses more consistent with the null hypothesis. Tables 5 and 6 in the Supplementary Materials use the PSB-adjusted



effect sizes model-averaged across all models, including models assuming the absence of a mean effect. These estimates, which are model-averaged also over the null, indicate stronger absolute bias compared to the conditional estimates presented in the main manuscript (Table 2).

Overestimation factor (OF) quantifies the degree to which the average effect sizes in meta-analyses overestimate the PSB-adjusted effect size estimates assuming the presence of the effect,

$$OF = \frac{\frac{1}{K} \sum_{k=1}^K \hat{y}_k}{\frac{1}{K} \sum_{k=1}^K \mu_{adj,k}}. \quad (4)$$

An overestimation factor larger than one is evidence of PSB. We use the delta method to obtain the confidence interval of the overestimation factor. In the Supplementary Materials, we also report medians and interquartile ranges of per meta-analysis overestimation factors,

$$OF_k = \frac{\hat{y}_k}{\mu_{adj,k}}. \quad (5)$$

However, note that  $OF_k$  can lead to non-sensible results as a meta-analysis with a positive mean effect and very small negative PSB-adjusted effect sizes estimate results in an extremely large negative  $OF_k$ .

### 2.3.3 | Evidence for publication selection bias

Posterior probability of PSB is an intuitive way of quantifying the evidence in favor of PSB. Similarly to the posterior probability of the presence of the effect, under the assumption of equal prior probability of the presence and the absence of PSB,  $p(\mathcal{H}_{PSB}) = p(\mathcal{H}_{NoPSB})$ , a posterior

probability larger than 0.5 indicates that the data are more likely under the presence of PSB. On the other hand, a posterior probability lower than 0.5 indicates that the data are more likely under the absence of PSB. As before, the Bayes factor for the presence of PSB,  $BF_{psb}$ ,<sup>1</sup> quantifies the change from prior to posterior odds for the presence of PSB. A Bayes factor in favor of the presence of PSB larger than one provides evidence in favor of the presence of PSB and lower than one provides evidence against the presence of PSB.

Relative publication probabilities quantify the relative probability of an estimate being published for a given  $p$ -value interval compared to estimates with statistically significant  $p$ -values. We use one-sided  $p$ -values, resulting in  $p$ -values larger than 0.5 corresponding to estimates in the opposite direction. To facilitate the interpretation we visualize a weight function that shows the change of relative publication probabilities across the range of  $p$ -values. We report the results only in Supplementary Materials.

Effect size inflation in imprecise estimates quantifies the relationship between the effect sizes and their standard errors. To facilitate the interpretation of the funnel asymmetry test, we visualize the bias in effect sizes as a function of standard errors (incorporating the quadratic term from the RoBMA model). We report the results only in Supplementary Materials.

## 3 | RESULTS

### 3.1 | Descriptives

Table 1 compares the characteristics of the meta-analyses from each field. Medical meta-analyses contain the smallest number of estimates per meta-analysis, followed by psychology and environmental sciences with five to six times the number of estimates compared to medicine. Finally, economics meta-analyses contain over 12 times the median number of estimates compared to medicine. Contrary to a naive expectation that more estimates may

TABLE 1 Summary of the data sets from each field.

Field	Meta-analyses	Estimates	Estimates/MA	Effect sizes ( $d$ )	Prop. significant
Medicine	67,386	597,699	5 (4, 10)	0.24 (0.09, 0.47)	0.39
Environmental	199	12,707	26 (11, 59)	0.62 (0.31, 0.95)	0.85
Psychology	605	23,563	18 (9, 40)	0.37 (0.18, 0.61)	0.78
Economics	327	91,421	66 (30, 283)	0.20 (0.09, 0.37)	0.82

Note: The number of estimates per meta-analysis (Estimates/MA) and the unweighted simple mean effect size of estimates within each meta-analysis (Effect Sizes) are reported as medians with the interquartile range (in parentheses). The proportion of the statistically significant (Prop. Significant) meta-analytic effect size estimates is based on a random-effect meta-analysis estimated via restricted maximum likelihood with  $\alpha = 0.05$  (removing one environmental sciences and 275 medical meta-analyses that did not converge).

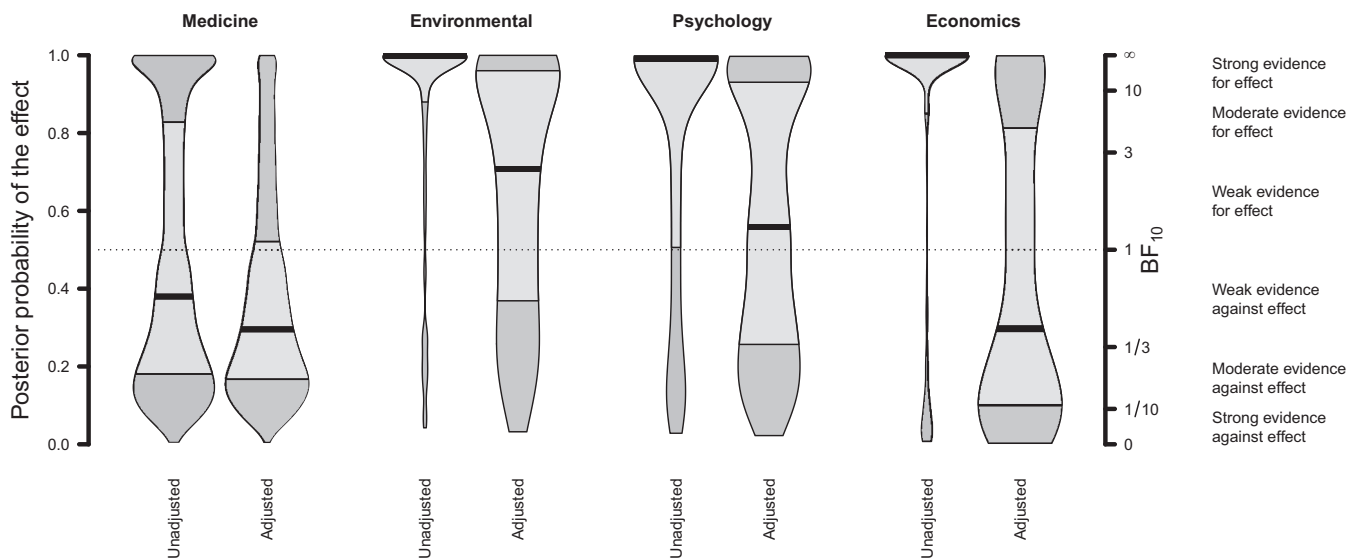
be conducted to establish smaller effects, economic and medical effect sizes are approximately the same magnitude (measured as a mean effect size per meta-analysis). Effect sizes in psychology are roughly twice as large as those in economics, and effect sizes in the environmental sciences are approximately three times larger than those in economics. Differences in the number of estimates per meta-analysis are most closely reflected in the proportion of statistically significant random-effects estimates. Notably, random-effects estimates in economics, psychology, and environmental sciences are statistically significant approximately twice as often as in medicine. This disparity in the proportion of statistically significant meta-analyses is consistent when comparing meta-analyses with a matched number of estimates across the disciplines, although the difference in mean effects is somewhat smaller (see Table 1 in the Supplementary Materials).

We summarize results from all meta-analyses, apart from seven medical meta-analyses that did not converge. See Supplementary Materials for analyses showing that matching meta-analyses based on the number of primary estimates within each meta-analysis does not meaningfully affect the conclusions.

### 3.2 | Evidence for the effect

Figure 1 shows medians, interquartile ranges, and distributions of the posterior probability of an effect before

and after adjusting for PSB. These distributions reveal several patterns. First, meta-analyses in economics, psychology, and environmental sciences predominantly show evidence for an effect before adjusting for PSB (unadjusted); whereas meta-analyses in medicine often display evidence against an effect. This disparity between the fields remains even when comparing meta-analyses with equal numbers of effect size estimates (see Supplementary Materials). After correcting for PSB, the posterior probability of an effect drops much more in economics, psychology, and environmental sciences (medians drop from 99.9% to 29.7%, from 98.9% to 55.7%, and from 99.8% to 70.7%, respectively) compared to medicine (38.0% to 29.7%). The pattern is especially striking in economics, where the median posterior probability of an effect drops by more than seventy percentage points after PSB correction. Mean decreases in posterior probabilities show a similar pattern but with somewhat smaller reductions (Tables 9 and 10 in Supplementary Materials). In all four disciplines, adjusting for PSB resulted in a substantial decrease in the strength of evidence for the effect: the proportion of meta-analyses with at least strong evidence for the presence of an effect (i.e.,  $BF_{10} > 10$ ) decreased from 20.2% to 5.3% in medicine, from 72.4% to 30.7% in environmental sciences, from 59.8% to 27.3% in psychology, and from 72.8% to 19.6% in economics. A comparable decrease was also present when comparing the proportion of meta-analyses with at least moderate evidence for the presence of an effect (i.e.,  $BF_{10} > 3$ ; from 28.9% to 12.3% in medicine,



**FIGURE 1** Median, interquartile range, and distribution of posterior probability for the presence of the effect before and after adjustment for publication selection bias in each field. The width of gray area indicates density, the light gray area indicates the interquartile range, and the black line indicates the median. The y-axis is scaled according to posterior probabilities assuming equal prior probabilities of presence versus absence of the effect. See the secondary y-axis for Bayes factors in favor of the effect that are independent of the assumed prior probability of the effect.

from 80.4% to 47.3% in environmental sciences, from 67.4% to 38.39% in psychology, and from 76.8% to 27.6% in economics).

Furthermore, we quantify the inflation of evidence in favor of an effect in meta-analyses via the evidence inflation factor—the increase in Bayes factor in favor of the effect due to the PSB. We find that meta-analyses in economics inflate the evidence by a median factor of 11,369, whereas the meta-analyses in environmental sciences and psychology inflate the evidence by “only” 45.9 and 30.0, respectively, and medicine by a median factor of 1.33. These extreme differences between the fields are largely driven by the disparity in the typical numbers of estimates per meta-analysis across the disciplines (Table 1). After standardizing the evidence inflation factor (sEIF) by the number of estimates per meta-analysis, we find that per estimate evidence inflation is the largest in psychology with a median factor of 1.27, followed by environmental sciences with a median factor of 1.22, economics with a median factor of 1.15, and medicine with a median factor of 1.05. Again, the strong evidence inflation in economics (11,369) is largely due to having many more estimates per meta-analysis than psychology and medicine. However, even after adjusting for different numbers, meta-analyses in medicine still show the least inflated evidence due to PSB.

### 3.3 | Effect size estimates

Table 2 summarizes the effect of PSB on effect sizes in each field. The first column reveals that environmental sciences, on average, suffer from as much as two and a half times larger absolute bias as medicine, economics, or psychology. The degree of absolute bias in environmental sciences is so large that it is comparable to average unadjusted effect sizes in other fields. Otherwise, medicine,

**TABLE 2** Summary of the footprints of publication selection bias on the meta-analytic effect sizes in the form of absolute bias (in Cohen's  $d$ ) and overestimation factor.

Field	Absolute bias ( $d$ )	Overestimation factor
Medicine	0.13 [0.12, 0.13]	1.62 [1.60, 1.64]
Environmental	0.33 [0.24, 0.41]	1.78 [1.42, 2.13]
Psychology	0.13 [0.11, 0.14]	1.39 [1.24, 1.55]
Economics	0.15 [0.13, 0.17]	2.16 [1.69, 2.64]

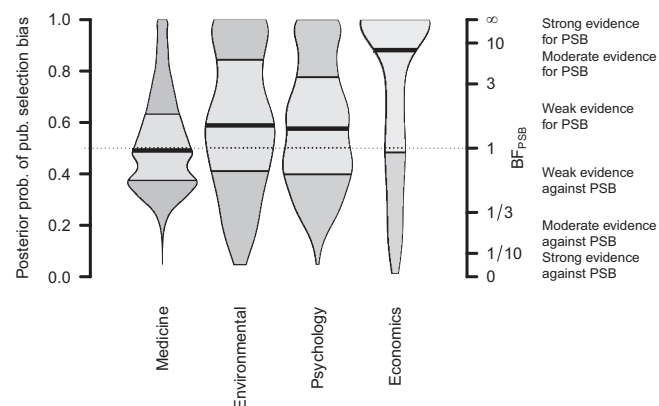
*Note:* The results are based on the comparison of publication bias adjusted meta-analytic effect size estimates assuming presence of the effect to the mean effect sizes per meta-analysis. The table displays means and 95% confidence intervals. (See Table 4 in the Supplementary Materials for medians and interquartile ranges.)

psychology, and economics share a comparable degree of absolute bias. The median absolute bias in each field is lower than the mean bias due to the right skew distribution of absolute biases (see Table 4 in the Supplementary Materials).

The second column of Table 2 displays the relative impact of PSB on meta-analytic estimates via the overestimation factor. On average, economics meta-analyses are, relatively, the most PSB exaggerated, inflating effect sizes by over two times; this corroborates a prior survey on power and bias.<sup>10</sup> Effect sizes in environmental sciences and medical meta-analyses show smaller yet notable relative effect size inflation. Finally, effect sizes in psychological meta-analyses are the least inflated with the average effect size exaggerated by 40%. In each field, the distribution of absolute biases is right-skewed; consequently, the per meta-analysis overestimation factor *median* is lower than the mean (see Table 4 in the Supplementary Materials). The median overestimation factor is relatively stable/decreasing with the increasing number of effect size estimates per meta-analysis, suggesting that the number of meta-analyses does not play a role in the relative size of PSB.

### 3.4 | Evidence for publication selection bias

Figure 2 shows medians, interquartile ranges, and distributions of the posterior probability of the PSB in each



**FIGURE 2** Median, interquartile range, and distribution of posterior probability for the presence of publication selection bias in each field. The width of gray area indicates density, the light gray area indicates the interquartile range, and the black line indicates the median. The  $y$ -axis is scaled according to posterior probabilities, assuming equal prior probabilities of presence versus absence of the publication selection bias. See the secondary  $y$ -axis for Bayes factors in favor of the publication selection bias that are independent of the assumed prior probability of the publication selection bias.



field. We find the most evidence for PSB in economics, where the typical evidence of the presence of publication bias is moderate, median  $BF_{PSB} = 7.27$  corresponding to 87.9% posterior probability of publication selection bias. Meta-analyses in environmental sciences and psychology have weak evidence in favor of PSB; even though the proportion of meta-analyses showing at least moderate evidence for PSB is still considerable (32.2% and 27.4%, respectively). Meta-analyses in medicine show the lowest proportion of at least moderate evidence in favor of PSB (12.9%). However, the proportion of meta-analyses consistent with the evidence of absence of PSB (i.e.,  $BF_{PSB} < 1/3$ ) is also the lowest in medicine (2.6%), indicating that the majority of medical meta-analyses is not informative enough to provide compelling evidence for or against publication bias. The proportion of meta-analyses with at least moderate evidence against PSB is somewhat higher in psychology (7.1%), economics (12.2%), and environmental sciences (12.6%). Meta-analyses with a larger number of effect size estimates present slightly more evidence in favor of the PSB; however, the overall disparity between the fields remains when comparing meta-analyses with a matched number of effect size estimates (Table 17 in Supplementary Materials).

#### 4 | CONCLUDING COMMENTS

We present a comprehensive assessment of publication selection bias and its effects on meta-analyses across medicine, environmental sciences, psychology, and economics. Novel methods and measures allowed us to quantify the evidence for the absence or presence of the mean effect and publication selection bias, as well as inflation of the evidence of the effect due to the publication selection bias. Furthermore, we estimated the bias and overestimation factor of the effect sizes of average estimates included in meta-analyses.

Our analysis is based on all effect size estimates found in these meta-analyses, regardless of the type of outcome or how they were analyzed. One can classify outcomes into three categories. First, some outcomes may have been pre-specified as being of primary interest to show a desirable effect (e.g., the effectiveness of a medication in reducing the risk of death). Second, some other outcomes are not pre-specified but may still be used to demonstrate some preferred outcome; thus, they may have larger analytical flexibility (e.g., using alternative measures of effectiveness) and thereby are potentially more affected by publication selection bias. Third, still other outcomes may have been collected and analyzed without any strong interest to show some significant result, or even

with some incentive to show non-significant results (e.g., outcomes on collected adverse events). Publication selection bias is expected mostly in the second category, while it may be less in the first category<sup>38</sup> and may be entirely absent in the third category.

Furthermore, we assumed independence of the reported primary estimates within and between meta-analyses; that is, each reported estimate is regarded to provide the same amount of new information as every other reported estimate. However, estimates may be dependent between meta-analyses, e.g., a single estimate might be used across multiple meta-analyses, and within meta-analyses, e.g., multiple estimates obtained from a single study/primary data set. As our data does not allow us to tackle those dependencies directly, we discuss how each independence violation might affect the results. The between meta-analysis dependency of estimates is of lesser importance as our inferences are concerned with the population of meta-analyses. Consequently, between meta-analysis dependency of estimates would only affect descriptive summaries of the estimates themselves. The within meta-analysis dependency of estimates is more problematic and can lead to (1) the overestimation of the strength of evidence, as the same primary data set is conditioned upon more than once, and (2) placing more weight on studies with multiple estimates. The first issue is partially mitigated via the standardized evidence inflation factor, which assesses the average evidence contribution of an estimate, that is, adjusting for the number of data sets conditioned upon. However, the absolute measures of evidence (i.e., evidence for the presence of the effect before and after publication bias adjustment or the evidence for publication bias) can be susceptible to overestimation, particularly in fields with relatively large within meta-analysis dependencies such as economics or environmental science (but see Supplementary Materials for comparison of meta-analyses with a matched number of effect size estimates). The second issue cannot be directly addressed; however, all presented measures are based on comparisons of two sets of models, both of which should be affected to a similar degree, thus hopefully canceling most of the bias that is generated by over-weighting studies with multiple estimates. Overall, we cannot exclude that the observed between-field differences may at least partially result from systematic differences in how meta-analyses themselves are conducted.

The milder publication selection bias in medical meta-analyses corroborates previous findings and might have multiple concurring explanations.<sup>9,16,19</sup> First, as in other disciplines, a large share of those medical meta-analyses with seemingly strong evidence no longer had strong evidence when PSB adjustment was made.

However, a much lower proportion of medical meta-analyses showed strong evidence of an effect compared to the other disciplines. Therefore, the difference between medicine and the other disciplines might be explained by the higher proportion of meta-analyses in medicine that showed weak evidence for an effect already before adjusting for publication selection bias. Second, medical studies may measure phenomena that are simpler and more stable, using methods that are more solidly and universally codified, which reduces researchers' "degrees of freedom" in generating and publishing evidence.<sup>9,16</sup> Third, it is also possible that the milder publication selection bias seen in medical meta-analyses is reflecting a larger share of meta-analyses that belonged to a category of outcomes with less pressure for publication selection bias. Finally, medical research makes wider use of research integrity practices, such as clinical trial registration, which might reduce the risk of publication selection bias.<sup>39</sup> Perhaps, medical research is, therefore, typically of a higher methodological quality and less subject to bias.<sup>9</sup>

In this paper, we documented the considerable impact of publication selection bias on meta-analyses in a variety of disciplines. Even though we can probe the footprint of these biases with the statistical techniques employed here, science ultimately needs to progress toward mitigating publication bias already while conducting and publishing the research. While the specific patterns of researchers' "degrees of freedom" and causes of publication selection bias are likely to vary widely across fields, our results suggest that the social sciences might especially benefit from adopting practices to mitigate these, including: pre-registration, greater transparency, and registered reports.<sup>40–42</sup>

## AUTHOR CONTRIBUTIONS

**František Bartoš:** Conceptualization; investigation; writing – original draft; writing – review and editing; visualization; validation; methodology; software; formal analysis; data curation. **Maximilian Maier:** Conceptualization; investigation; writing – original draft; methodology; validation; writing – review and editing; visualization; software; formal analysis; data curation. **Eric-Jan Wagenmakers:** Supervision; resources; methodology; writing – review and editing; writing – original draft; conceptualization; funding acquisition. **Franziska Nippold:** Investigation; writing – review and editing; resources. **Hristos Doucouliagos:** Investigation; writing – review and editing; resources. **John Ioannidis:** Investigation; writing – review and editing; resources. **Willem M. Otte:** Resources; writing – review and editing; investigation. **Martina Sladekova:** Resources; writing – review and editing; investigation. **Teshome K**

**Deressa:** Investigation; writing – review and editing; resources. **Stephan B Bruns:** Investigation; writing – review and editing; resources. **Daniele Fanelli:** Investigation; writing – review and editing; resources. **T.D. Stanley:** Supervision; resources; writing – review and editing; conceptualization; investigation; writing – original draft; methodology; project administration.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The majority of the data is available from the references resources (with a portion of the raw data unavailable due to restrictions from the original owners). Anonymized summary estimates for reproducing the analyses is accessible at OSF. See <https://osf.io/bgfzp/> for data and analysis scripts.

## ORCID

František Bartoš  <https://orcid.org/0000-0002-0018-5573>


Maximilian Maier  <https://orcid.org/0000-0002-9873-6096>

Eric-Jan Wagenmakers  <https://orcid.org/0000-0003-1596-1034>

Hristos Doucouliagos  <https://orcid.org/0000-0001-5269-3556>

John P. A. Ioannidis  <https://orcid.org/0000-0003-3118-6859>

Willem M. Otte  <https://orcid.org/0000-0003-1511-6834>

Martina Sladekova  <https://orcid.org/0000-0001-5059-6576>

Teshome K. Deressa  <https://orcid.org/0000-0003-1351-1849>

Stephan B. Bruns  <https://orcid.org/0000-0002-3028-9699>

Daniele Fanelli  <https://orcid.org/0000-0003-1780-1958>

T. D. Stanley  <https://orcid.org/0000-0002-3205-1983>

## ENDNOTE

<sup>1</sup> In other publications, we abbreviate this as BF<sub>pb</sub> or BF<sub>ow</sub>.

## REFERENCES

- Chavalarias D, Ioannidis JP. Science mapping analysis characterizes 235 biases in biomedical research. *J Clin Epidemiol*. 2010;63(11):1205-1215.
- Dwan K, Altman DG, Arnaiz JA, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One*. 2008;3(8):e3081.
- Rosenthal R, Gaito J. Further evidence for the cliff effect in interpretation of levels of significance. *Psychol Rep*. 1964;15(2):570.
- Wicherts JM. The weak spots in contemporary science (and how to fix them). *Animals*. 2017;7(12):90-119.

5. Otte WM, Vinkers CH, Habets PC, IJzendoorn vDG, Tijdkink JK. Analysis of 567,758 randomized controlled trials published over 30 years reveals trends in phrases used to discuss results that do not reach statistical significance. *PLoS Biol.* 2022;20(2):e3001562.
6. John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol Sci.* 2012;23(5):524-532.
7. Fiedler K, Schwarz N. Questionable research practices revisited. *Soc Psychol Personal Sci.* 2016;7(1):45-52.
8. De Winter JC, Dodou D. A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ.* 2015;3:e733.
9. Fanelli D, Costas R, Ioannidis JP. Meta-assessment of bias in science. *Proc Natl Acad Sci.* 2017;114(14):3714-3719.
10. Ioannidis JP, Stanley T, Doucouliagos H. The power of bias in economics research. *Econ J.* 2017;127(605):F236-F265.
11. Mathur VW. Finding common ground in meta-analysis “wars” on violent video games. *Perspect Psychol Sci.* 2019;14(4):705-708.
12. Stanley TD, Carter EC, Doucouliagos H. What meta-analyses reveal about the replicability of psychological research. *Psychol Bull.* 2018;144(12):1325-1346.
13. Van Aert RC, Wicherts JM, Van Assen MA. Publication bias examined in meta-analyses from psychology and medicine: a meta-meta-analysis. *PLoS One.* 2019;14(4):e0215052.
14. Schwab S, Kreiliger G, Held L. Assessing treatment effects and publication bias across different specialties in medicine: a meta-epidemiological study. *BMJ Open.* 2021;11(9):e045942.
15. Kühberger A, Fritz A, Scherndl T. Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PLoS One.* 2014;9(9):e105825.
16. Fanelli D. “Positive” results increase down the hierarchy of the sciences. *PLoS One.* 2010;5(4):e10068.
17. Ioannidis JP. Excess significance bias in the literature on brain volume abnormalities. *Arch Gen Psychiatry.* 2011;68(8):773-780.
18. Scheel AM, Schijen MR, Lakens D. An excess of positive results: comparing the standard psychology literature with registered reports. *Adv Methods Pract Psychol Sci.* 2021;4(2):1-12.
19. 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.
20. Fanelli D. Negative results are disappearing from most disciplines and countries. *Scientometrics.* 2012;90(3):891-904.
21. Deressa TK, Stern DI, Vangronsveld J, et al. More than half of statistically significant research findings in the environmental sciences are actually not. Submitted for publication 2022.
22. Sladekova M, Webb LE, Field AP. Estimating the change in meta-analytic effect size estimates after the application of publication bias adjustment methods. *Psychol Methods.* 2022;28:664-686.
23. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw.* 2010;36(3):1-48.
24. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. *Introduction to Meta-Analysis.* John Wiley & Sons; 2009.
25. Maier M, Bartoš F, Wagenmakers EJ. Robust Bayesian meta-analysis: addressing publication bias with model-averaging. *Psychol Methods.* 2022;28:107-122.
26. Bartoš F, Maier M, Wagenmakers EJ, Doucouliagos H, Stanley TD. Robust Bayesian meta-analysis: model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods.* 2022;14(1):99-116.
27. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Stat Sci.* 1999;14(4):382-401.
28. Fragoso TM, Bertoli W, Louzada F. Bayesian model averaging: a systematic review and conceptual classification. *Int Stat Rev.* 2018;86(1):1-28.
29. Vevea JL, Hedges LV. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika.* 1995;60(3):419-435.
30. Stanley TD, Doucouliagos H, Ioannidis JP. Finding the power to reduce publication bias. *Stat Med.* 2017;36(10):1580-1598.
31. Erp VS, Verhagen J, Grasman RP, Wagenmakers EJ. Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990–2013. *J Open Psychol Data.* 2017;5(1):4.
32. Wrinch D, Jeffreys H. On certain fundamental principles of scientific inquiry. *Phil Mag.* 1921;42:369-390.
33. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc.* 1995;90(430):773-795.
34. Jeffreys H. *Theory of Probability.* 1st ed. Oxford University Press; 1939.
35. Lee MD, Wagenmakers EJ. *Bayesian Cognitive Modeling: A Practical Course.* Cambridge University Press; 2013.
36. Wagenmakers EJ, Morey RD, Lee MD. Bayesian benefits for the pragmatic researcher. *Curr Dir Psychol Sci.* 2016;25(3):169-176.
37. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat.* 2016;70(2):129-133.
38. Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. *Jama.* 1998;279(14):1089-1093.
39. Laine C, Horton R, DeAngelis CD, et al. Clinical trial registration—looking back and moving ahead. *N Engl J Med.* 2007;356(26):2734-2736.
40. Chambers CD. Registered reports: a new publishing initiative at cortex. *Cortex.* 2013;49(3):609-610.
41. Chambers CD, Dienes Z, McIntosh RD, Rotshtein P, Willmes K. Registered reports: realigning incentives in scientific publishing. *Cortex.* 2015;66:A1-A2.
42. Akker v dOR, Weston S, Campbell L, et al. Preregistration of secondary data analysis: a template and tutorial. *Meta-Psychology.* 2021;5:5.

## AUTHOR BIOGRAPHIES

**František Bartoš** is a PhD candidate at the Department of Psychological Methods at the University of Amsterdam.

**Maximilian Maier** is a PhD candidate in Experimental Psychology at University College London.

**Eric-Jan Wagenmakers** is professor of Bayesian Methodology at the Department of Psychological Methods at the University of Amsterdam.

**Franziska Nippold** is a former research masters student of the Psychology department at the University of Amsterdam.

**Hristos Doucouliagos** is Emeritus Professor of Economics at Deakin University, Melbourne Australia.

**John P. A. Ioannidis** is Professor of Medicine, of Epidemiology and Population Health, and (by courtesy) of Biomedical Data Science, and of Statistics at Stanford University where he leads the Meta-Research Innovation Center at Stanford (METRICS).

**Willem M. Otte** is Associate Professor at Utrecht University and University Medical Center Utrecht, working on methods for clinical epidemiology.

**Martina Sladekova** is a Lecturer in Psychology and a PhD Candidate at the School of Psychology, University of Sussex.

**Teshome K. Deressa** is a PhD candidate at the Centre for Environmental Sciences at Hasselt University.

**Stephan B. Bruns** is Associate Professor of Environmental Economics at Hasselt University.

**Daniele Fanelli** is an Assistant Professor at the Doctoral Centre in the School of Social Sciences, Heriot-Watt University, and a Visiting Fellow at the

Department of Methodology, London School of Economics and Political Science.

**T. D. Stanley** is Professor of Meta-Analysis, Emeritus, and Honorary Professor of Economics at Deakin University, Melbourne Australia. He is an elected Fellow of the Society of Research Synthesis Methods and Convener of the Meta-Analysis of Economics Research Network and was the Julia Mobley Professor of Economics at Hendrix College, Conway USA.

## 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:** Bartoš F, Maier M, Wagenmakers E-J, et al. Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics. *Res Syn Meth.* 2024;1-12. doi:[10.1002/jrsm.1703](https://doi.org/10.1002/jrsm.1703)