More Than Half of Statistically Significant Research Findings in the Environmental Sciences are Actually Not

Teshome K. Deressa*[a], David I. Stern[b], Jaco Vangronsveld[a], Jan Minx[c,d], Sebastien Lizin[a], Robert Malina[a,e], Stephan B. Bruns†[a,f]

[a] *Centre for Environmental Sciences, Hasselt University, Martelarenlaan 42, 3500 Hasselt, Belgium*

[b] *Arndt-Corden Department of Economics, Crawford School of Public Policy, The Australian National University, Australia*

[c] *Mercator Research Institute on Global Commons and Climate Change, Berlin, Germany*

[d] *Priestley International Centre on Climate, School of Earth and Environment, University of Leeds, United Kingdom*

[e] *Department of Aeronautics and Astronautics, Laboratory for Aviation and the Environment, Massachusetts Institute of Technology, USA*

[f] *Department of Economics, University of Göttingen, Germany*

Abstract

Researchers have incentives to search for and selectively report findings that appear to be statistically significant and/or conform to prior beliefs. Such selective reporting practices, including p-hacking and publication bias, can lead to a distorted set of results being published, potentially undermining the process of knowledge accumulation and evidence-based decision making. We take stock of the state of empirical research in the environmental sciences using 67,947 statistical tests obtained from 547 meta-analyses. We find that 59% of the p-values that were reported as significant are not actually expected to be statistically significant. The median power of these tests is between 6% to 12%, which is the lowest yet identified for any discipline. Only 8% of tests are adequately powered with statistical power of 80% or more. Exploratory regressions suggest that increased statistical power and the use of experimental research designs reduce the extent of selective reporting. Differences between subfields can be mostly explained by methodological differences. To improve the environmental sciences evidence base, researchers should pay more attention to statistical power, but incentives for selective reporting may remain even with adequate statistical power. Ultimately, a paradigm shift towards open science is needed to ensure the reliability of published empirical research.

Introduction

In recent years, the number of empirical studies in the environmental sciences has increased substantially (Zhong et al. 2021). However, recent studies indicate that much published research in other empirical disciplines cannot be replicated, which questions the reliability of published research (Open Science Collaboration 2015; Camerer et al. 2016; Chang and Li 2022). This poses a fundamental challenge for cumulative knowledge production (Furman et al. 2012) and evidence-based decision- and policy-making (Chen et al. 2012). The main reason for this lack of reliability is researchers' incentive to selectively report statistically significant findings (Ioannidis 2005). Small sample sizes and low statistical power are particularly challenging, as they reduce the chance of finding a true effect, as well as the likelihood that a statistically significant result reflects a true effect (Button et al. 2013).

In this paper, we take stock of the state of empirical research in the environmental sciences: (1) we estimate the extent of selectively reported p-values for the entire environmental sciences and for seven of its subfields; (2) we provide the first large-scale assessment of statistical power in this domain using a total number of

*. teshome.deressa@uhasselt.be
†. stephan.bruns@uhasselt.be

67,947 statistical tests obtained from a stratified sample of 547 meta-analyses; and (3) we explore potential determinants of selective reporting using regression analysis.

We find that 59% of reported significant p-values should have not been reported as statistically significant. We also find a retrospective median statistical power (Button et al. 2013; Stanley et al. 2022) of 6% to 12%, which is approximately half the median power found in neuroscience, economics, and biomedical research. Our exploratory analysis reveals that the extent of selective reporting can be reduced by increasing statistical power and, to some extent, by using experimental research designs.

Our analysis suggests that low statistical power and the corresponding large extent of selective reporting undermine the reliability of published research in the environmental sciences. The environmental sciences can benefit from measures taken in other disciplines that cope with similar challenges by pre-registering study protocols and encouraging open science, including data and code availability policies and incentivizing replication studies. Statistical power should be routinely considered and reported in order to reduce the risk of p-hacking and the resulting exaggerated effect sizes due to low sample sizes.

p-hacking undermines the reliability of reported *p*-values

Researchers need to make a multitude of choices, which may affect the outcomes of their study (Simmons et al. 2011). When different research teams analyze the same research question using the same data set, considerable heterogeneity in both study designs and outcomes can be observed (Botvinik-Nezer et al. 2020; Breznau et al. 2022; Huntington-Klein et al. 2021; Silberzahn et al. 2018). As researchers face incentives to search for and to report statistically significant findings – particularly those that conform with prior beliefs or sensational findings and large effect sizes (Ioannidis 2005) – they may consciously or unconsciously use the flexibility in possible research designs to obtain desired study outcomes (Bruns and Kalthaus 2020). Published p-values may then be neither replicable nor robust, nor convey reliable information about whether the null hypothesis of no effect should be rejected.

For example, a researcher may run an experiment with multiple outcome variables. Even if the treatment has no true effect on any of the outcome variables, the rate of false-positive findings increases with the number of each outcome variable. For ten independent outcome variables, the probability of finding a statistically significant effect increases from 5% to 40%, even in the absence of any true effect. The analysis of observational data without randomization usually provides even more leeway in specifying the study design (Bruns and Kalthaus 2020; Leamer 1983). Once the outcomes are known, researchers may (unconsciously) convince themselves that the study design with the desirable outcome is the optimal design. Behavioral biases, such as motivated reasoning, and a lack of statistical training mean that p-hacking is often an unconscious process rather than a deliberate search for statistically significant findings (Bruns and Kalthaus 2020; Touchon and McCoy 2016; Bastardi et al. 2011; Nickerson 1998).

While p-hacking may occur within studies, entire studies may be suppressed by editors, reviewers, or the authors themselves, if, for example, publication is deemed unlikely due to a lack of statistically significant findings. Such selective reporting at the study level is known as the file-drawer problem or publication bias (Franco et al. 2014; Rosenthal 1979). Several studies have diagnosed the presence of selective reporting in multiple scientific fields, including economics (Brodeur et al. 2020; Ioannidis et al. 2017), management (Harrison et al. 2017; Baum and Bromiley 2019; Rost and Ehrmann 2017), innovation research (Bruns et al. 2019), impact evaluation studies (Vivalt 2019), political science and sociology (Gerber and Malhotra, 2008a, 2008b), and the biomedical sciences (Turner et al. 2008; Albarqouni et al. 2017).

Low statistical power amplifies *p*-hacking

Statistical power is the probability that a test detects the effect if the effect does indeed exist. Low statistical power reduces the chance that a statistically significant result reflects a true effect and, as such, facilitates *p*-hacking (Button et al. 2013). Median statistical power has been shown to be very low (18–21%) in fields such as neuroscience, biomedical research, and economics (Button et al. 2013; Ioannidis et al. 2017; Dumas-Mallet et al. 2017; Lamberink et al. 2018) and higher in fields such as psychological research (36%), intelligence research (52%), and criminology (71%) (Stanley et al. 2018; Nuijten et al. 2020; Barnes et al. 2020). A common threshold for adequate power is 80%, that is, repeated sampling results in detection of the effect 80% of the time if the analyzed effect exists (Button et al. 2013). Power increases with the size of both the true effect and the sample size. While the former is beyond the control of the researcher, sample size is more likely to be at least partially in the researcher's control. Budget constraints in experimental research are a challenge, but might be overcome by funding agencies demanding adequately powered research. Similarly, advances in data gathering methods – such as web scraping and text mining – can help to increase sample sizes in observational research. If researchers do not base sample sizes on power considerations, sample sizes might be frequently too small to reliably detect a true effect. In such cases, *p*-hacking is typically used to exaggerate the estimated effect size, as a large effect estimate may ensure statistical significance. This calls for caution when interpreting published effect sizes, especially in cases where these are used to inform policy and decision-making.

Statistical power has been shown to be low in ecology – one of the subfields of the environmental sciences analyzed in this paper. A recent meta-analysis of 553 statistical tests from 250 animal tracking studies found a median power of 9% for small, 31% for medium, and 65% for large genuine effects using Cohen's classification (Cleasby et al. 2021). In another analysis of 278 statistical tests, median power was 7–8% for small effects and 23–26% for medium effects, respectively (Smith et al. 2011).

Results

Characteristics of the data

Summary statistics for the sampled meta-analyses and primary estimates are given in Table 1. From 547 meta-analyses, we obtained a total of 67,947 primary estimates along with their respective standard errors. The average number of primary estimates per meta-analysis is 124, with a range of 5 to 4089. The majority of meta-analyses (389 of 547) synthesized observational studies. About 46% of the meta-analyses (252) have no adequately powered estimates (power of 80% or higher). The proportion of meta-analyses that follow standard reporting guidelines such as "Preferred Reporting Items for Systematic Reviews and Meta-Analyses" (PRISMA) (Moher et al. 2009) or "RepOrting standards for Systematic Evidence Syntheses" (ROSES) (Haddaway et al. 2018) is 47% (258), and the proportion of meta-analyses that pre-registered their study protocols in platforms like Prospective Register for Systematic Reviews (PROSPERO) is 6.6% (36).

Selective reporting

In Figure 1, we compare reported z-values from our data set with counterfactual z-values, assuming the absence of any selective reporting in the research and publication process, in order to estimate the extent of selectively reported statistical findings. Specifically, we use each statistical test's reported standard error, but replace the estimated effect with our own estimate of the respective true effect. This true effect is approximated by the meta-average of the respective meta-analysis (Bruns et al. 2022). The reported and counterfactual distribution of z-values intersect at the common threshold of statistical significance $\alpha = 0.05$ ($z = 1.96$). To the left of the threshold (for $z < 1.96$), there are fewer than expected z-values. To the right of the threshold, there are more than the expected number of z-values. This suggests that researchers use this

Table 1: Characteristics of included meta-analyses and corresponding primary estimates.

	No. of meta-analyses	No. of primary estimates	Primary estimates per meta-analyses					
			Mean	Min	Q25	Q50	Q75	Max
All meta-analyses	547	67947	124	5	15	39	93	4089
Research design								
Observational	389	40230	103	5	12	29	64	4089
Experimental	158	27717	175	5	35	78	163	2687
Statistical power								
SAPE >0	295	36059	122	5	13	32	78	4089
SAPE = 0	252	31888	127	5	17	52	110	2687
Followed guideline								
Yes	258	22873	89	5	11	26	70	2203
No	289	45074	156	5	21	51	114	4089
Protocol pre-registered								
Yes	36	2871	80	5	14	40	62	824
No	511	65076	127	5	15	39	97	4089

Notes: 'Observational' denotes meta-analyses of observational primary studies or a mixture of observational and experimental primary studies, whereas 'Experimental' denotes meta-analyses of exclusively experimental primary studies. 'SAPE' denotes the share of adequately powered estimates (see *Methods* section). 'Followed guideline' denotes the use of guidelines in conducting the meta-analysis and 'Protocol pre-registered' denotes the use of a pre-analysis plan specified before conducting the meta-analysis. Q25, Q50, and Q75 refer to the quantiles of the primary estimates per meta-analysis.

threshold, either consciously or unconsciously, to selectively report results with systematic overreporting of significant and underreporting of non-significant results.

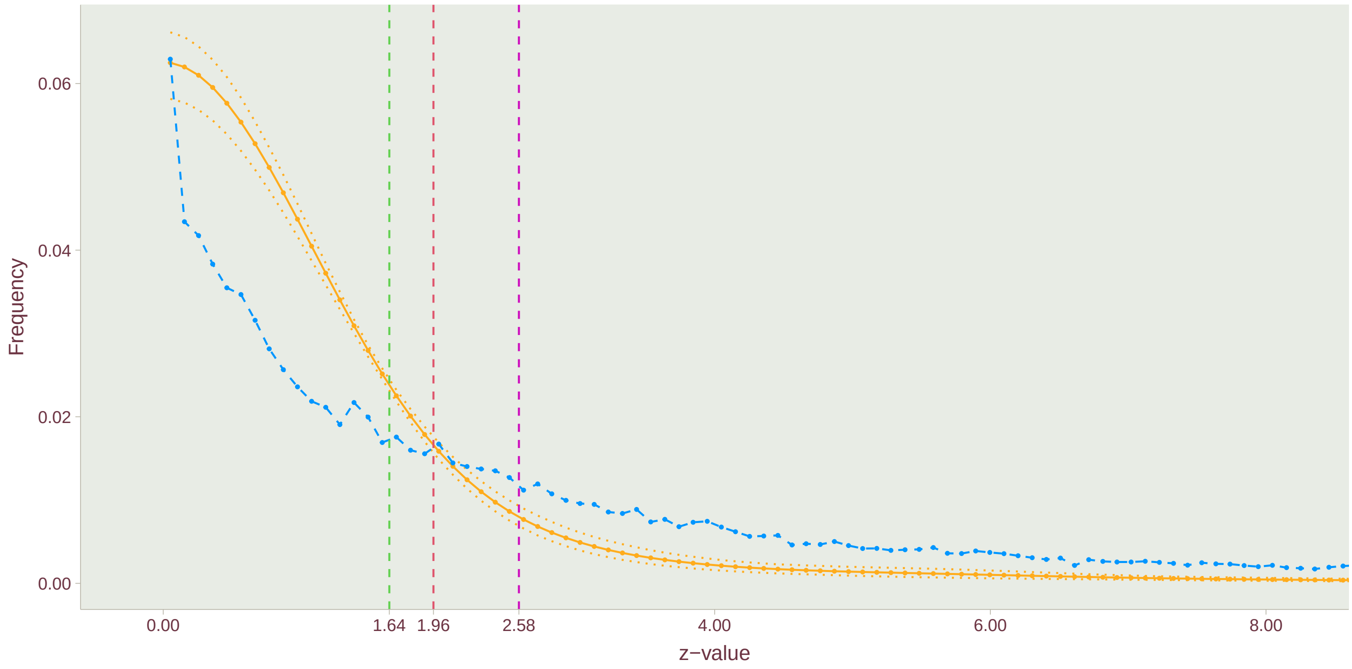


Figure 1: Distributions of reported (blue, dashed line) and counterfactual (orange, solid line) z-values with dotted 95% confidence intervals based on bootstrapping clustered by meta-studies. The dots are placed at the center of each interval using a grid of 0.1025. The dashed vertical lines mark the critical values for $\alpha = 0.1$ $(z = 1.64), \alpha = 0.05$ $(z = 1.96)$ and $\alpha = 0.01$ $(z = 2.58)$.

Using bins of p-values, we quantify this overreporting and underreporting as a share of all p-values in Table 2. There is a lack of non-significant p-values $(p > 0.05)$. Specifically, in most p-value bins, about 3-4% of the total probability mass is missing, whereas there is an abundance of statistically significant p-values with a large overrepresentation of p-values that are significant at the 0.001 level (22.6%).

The extent of selective reporting amounts to 27% of all p-values (ESR_{all}) using the $\alpha = 0.05$ threshold for statistical significance. The extent of selective reporting rises to 59% expressed as a share of only the statistically significant p-values ($ESR_{sig.}$). In other words, we expect, in the absence of any selective reporting in the publication and research process, that 59% fewer p-values should be statistically significant at the $\alpha = 0.05$ level than were reported.

Table 2: Extent of selective reporting.

p-value interval	Difference between observed and counterfactual p-values	95% CI
$0.9 < p$	-0.004	[-0.008, 0.002]
$0.8 < p < 0.9$	-0.023	[-0.028, -0.018]
$0.7 < p < 0.8$	-0.027	[-0.031, -0.022]
$0.6 < p < 0.7$	-0.030	[-0.034, -0.025]
$0.5 < p < 0.6$	-0.031	[-0.035, -0.026]
$0.4 < p < 0.5$	-0.035	[-0.039, -0.030]
$0.3 < p < 0.4$	-0.039	[-0.043, -0.035]
$0.2 < p < 0.3$	-0.040	[-0.043, -0.036]
$0.1 < p < 0.2$	-0.033	[-0.036, -0.030]
$0.05 < p < 0.1$	-0.011	[-0.014, -0.009]
$0.01 < p < 0.05$	0.014	[0.007, 0.021]
$0.001 < p < 0.01$	0.032	[0.024, 0.039]
$p < 0.001$	0.226	[0.199, 0.249]
ESR_{all}	0.273	[0.235, 0.307]
$ESR_{sig.}$	0.589	[0.496, 0.663]
No. of meta-analyses	547	
No. of tests	67947	

Notes: Differences between observed p-values and counterfactually expected p-values per p-value bins as a share of all p-values are reported. Negative signs indicate that p-values are missing in the respective bin compared to what is expected in the absence of any biases. ESR_{all} represents the estimated extent of selective reporting as a share of *all* p-values, whereas $ESR_{sig.}$ represents the estimated extent of selective reporting as a share of *only* significant p-values. 95% confidence intervals based on bootstrapping clustered at the level of meta-studies.

Limiting our analysis to subfields with 20 or more meta-analyses (see Tables S5–S7 and Figure S2 in the Online Appendix), the extent of selective reporting expressed as a share of the significant p-values is approximately 50% in ‘Ecology’, ‘Environmental Chemistry’ and ‘Health, Toxicology and Mutagenesis’. The ‘Nature and Landscape Conservation’ subfield stands out with an estimated extent of selective reporting of 77%. For the subfields with less than 20 meta-analyses, the extent of selective reporting is 80% for ‘Environmental Engineering’, 23% for ‘Management, Monitoring, Policy and Law’, and 14% for ‘Water Science and Technology’.

We test the robustness of our main estimate of the extent of selective reporting ($ESR_{sig.} = 0.59$) to various assumptions made in the analysis (Tables S1–S4 and Figure S1 in the Online Appendix). We consider the main estimate to be robust if the confidence intervals overlap. The estimate is robust with respect to the use of alternative approaches to approximate the true effect. Using half the meta-average to account for the inflation of meta-averages in the presence of selective reporting (Bruns et al. 2022), results in an estimate of $ESR_{sig.} = 0.74$ and the confidence intervals no longer overlap. Assuming two or three true effects per meta-analysis, rather than a single true effect, results in estimates of the extent of selective reporting of 53% and 52%, respectively. These estimates are also within the confidence interval of our main estimate. The estimate of the extent of selective reporting rises to 60% and 75% when the standard errors of the estimates are multiplied by 1.5 and 2, respectively. In the latter case, the confidence intervals are not overlapping. The rationale for using larger standard errors is that the standard errors themselves might be biased downwards if they are subjected to selective reporting. Including very large z-values in the analysis is likely to result in overestimating the genuine effect and, thus, a lower estimate of selective reporting. Including z-values up to 50 ($n = 71,921$) results in an estimate of $ESR_{sig.} = 0.43$ and up to 100 ($n = 73,499$) in an estimate of $ESR_{sig.} = 0.32$, and in the latter case the confidence intervals are not overlapping.

Statistical power

The levels of statistical power in the environmental sciences and seven of its subfields are reported in Table 3 using the standard errors of each statistical test in our data set and approximating the true effect for each meta-analysis using its meta-averages. Previous studies for other disciplines report either the median statistical power using all primary estimates (Nuijten et al. 2020; Dumas-Mallet et al. 2017; Pietschnig et al. 2019) or the median of medians by first calculating the median statistical power per meta-analysis (Ioannidis et al. 2017; Stanley et al. 2018). Both measures indicate that power is low in the environmental sciences and ranges from 5.5% to 55% across the subfields. The mean and quantiles indicate that the statistical power distributions are right-skewed. The share of adequately-powered estimates ranges between 1.3% and 28% across subfields. For the environmental sciences in total, only 8.2% of all primary estimates are adequately powered. Overall, the median power of empirical studies in environmental sciences is 6% to 12% depending on the definition of median power. This is lower than the median power of all other scientific disciplines that have been investigated so far (ranging from 16% for randomized clinical trials to 71% in criminology). From the subfields with 20 or more meta-analyses, 'Health, Toxicology and Mutagenesis' has the highest power, while the subfield 'Nature and Landscape Conservation' has the lowest.

Table 3: Statistical power in environmental sciences.

Subfield	No. of meta-analyses	Median of medians	Median	Mean	Q25	Q75	SAPE
All meta-analyses	547	0.123	0.060	0.190	0.051	0.147	0.082
Ecology	163	0.103	0.056	0.196	0.050	0.159	0.085
Environmental Chemistry	72	0.079	0.085	0.209	0.051	0.194	0.088
Environmental Engineering	11	0.084	0.055	0.105	0.051	0.067	0.037
Health, Toxicology and Mutagenesis	196	0.388	0.063	0.263	0.050	0.159	0.169
Management, Monitoring, Policy and	19	0.190	0.172	0.341	0.057	0.651	0.158
Nature and Landscape Conservation	71	0.064	0.058	0.110	0.052	0.085	0.013
Water Science and Technology	15	0.551	0.444	0.499	0.172	0.875	0.278

Notes: Median of medians denotes the median of the medians of statistical power per meta-analysis. Whereas the columns Median, Mean, Q25, Q75, and SAPE (Share of Adequately Powered Estimates) are directly based on the respective distribution of primary estimates.

Figure 2(a) shows the distribution of the median power of the primary estimates for each meta-analysis. More than half (59%) of the meta-analyses have a median power of 20% or less, while only about a fifth (22%) have a median power larger than 80%. The U-shape pattern of the histogram looks similar to those reported in other fields, such as economics (Ioannidis et al. 2017). 71% of the meta-analyses with a median power of more than 80% stem from the subfield of 'Health, Toxicology and Mutagenesis'. As depicted in Figure 2(b), 252 out of 547 meta-analyses (46%) do not have any adequately powered primary estimates and only 21% of meta-analyses have a share of adequately powered estimates that is larger than 20%.

Explaining variation in selective reporting

Potential determinants of selective reporting are explored in Table 4. An increase of one percentage point in the share of adequately powered studies reduces the extent of selective reporting by about 0.7%. This finding is expected, as low power may incentivize researchers to consciously or unconsciously search for significant effects (Button et al. 2013). Moreover, we use the publication year of the meta-analysis to identify potential

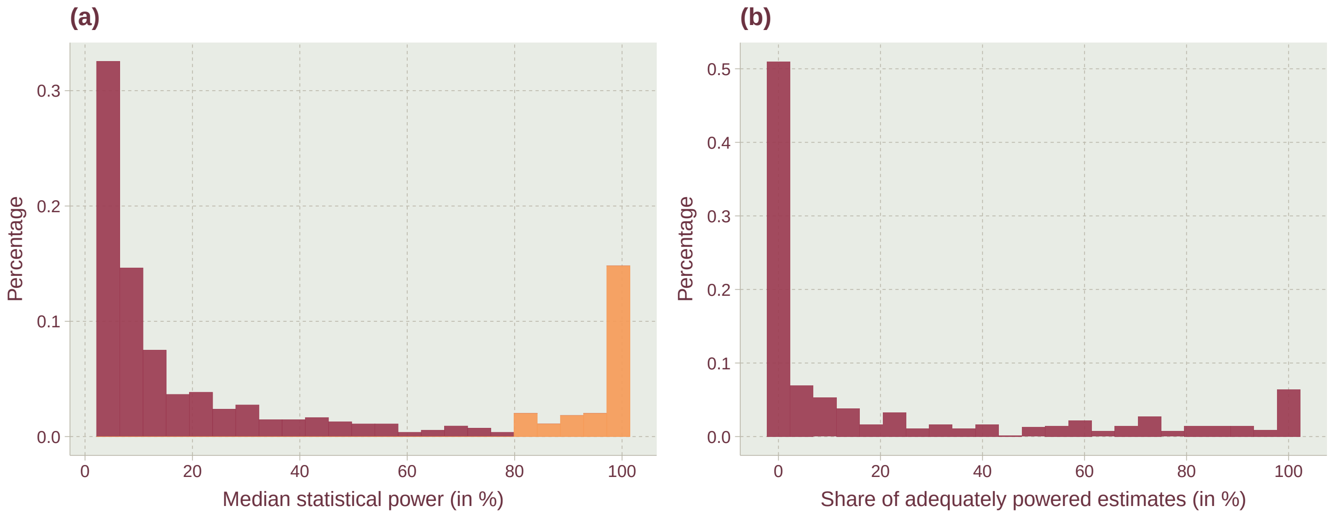


Figure 2: (a) Distribution of median power of primary estimates per meta-analysis obtained from 547 meta-analyses. Adequately powered estimates are highlighted in orange. (b) Distribution of the share of adequately powered estimates in 547 meta-analyses.

time trends. This assumes that more recently published meta-analyses cover more recently published primary studies. The extent of selective reporting decreases with the publication year of the meta-analysis.

Inclusion of dummies for subfields increases the adjusted R^2 by 1% (Model 2 compared to Model 1), indicating that there is significant variation across subfields. Two subfields show less selective reporting compared to the reference group 'Health, Toxicology and Mutagenesis', but these two subfields have fewer than 20 meta-analyses each. These findings indicate that differences between subfields can be mostly explained by differences in statistical power.

We also explore the effect of an interaction between the share of adequately powered studies and experimental research designs (Model 3) because the unconditional correlation between these two variables of -0.24 shows that experimental studies are more underpowered than observational studies. The interaction effect helps to disentangle the effect of experimental research designs and statistical power, showing that research fields with experimental research designs have less selective reporting if they have the same power as observational studies. The interaction effect and the effect of time are sensitive to the inclusion of outliers (see *Methods* section).

Finally, we also explore whether pre-registering a protocol or following reporting guidelines for the meta-analysis is related to the extent of selective reporting in the primary studies. For example, a meta-analysis of a field with contesting views and corresponding selective reporting may follow established reporting guidelines and use pre-registration to improve its reliability. We do not find evidence for such associations for meta-analyses, but it is more likely that these associations are present at the level of primary studies.

Discussion

This study takes stock of the state of empirical research in the environmental sciences using a stratified random sample of 547 meta-analyses with 67,947 primary estimates. We find that published empirical studies in the environmental sciences are severely underpowered with a median statistical power of 6% to 12% and only 8.2% of all estimates being adequately powered. Low statistical power begets selective reporting, and we find, correspondingly, that 59% of reported significant p-values should have not been published as

Table 4: Regression results using OLS (Dependent variable: $ESR_{sig.}$).

	Model 1	Model 2	Model 3
Intercept	41.729***	40.379***	39.941***
	(15.604)	(15.340)	(15.177)
Share of adequately powered estimates	-0.007***	-0.007***	-0.007***
	(0.001)	(0.001)	(0.001)
Experimental	-0.015	-0.006	0.039
	(0.056)	(0.055)	(0.062)
Share of adequately powered estimates*Experimental			-0.004**
			(0.002)
Followed reporting guideline (yes/no)	0.067	0.050	0.052
	(0.045)	(0.042)	(0.042)
Protocol pre-registered (yes/no)	0.084	0.084	0.080
	(0.106)	(0.090)	(0.091)
Log total number of p-values	0.036*	0.034	0.029
	(0.021)	(0.021)	(0.022)
Log number of independent studies included in meta-analysis	0.002	0.012	0.016
	(0.033)	(0.033)	(0.033)
Publication year	-0.021***	-0.020***	-0.020***
	(0.008)	(0.008)	(0.008)
Ecology		-0.085	-0.073
		(0.059)	(0.059)
Environmental Chemistry		-0.010	-0.010
		(0.070)	(0.070)
Environmental Engineering		0.045	0.031
		(0.128)	(0.130)
Management, Monitoring, Policy and Law		-0.123*	-0.095
		(0.068)	(0.061)
Nature and Landscape Conservation		-0.058	-0.048
		(0.065)	(0.064)
Water Science and Technology		-0.226**	-0.219**
		(0.110)	(0.109)
Adjusted R^2	0.395	0.405	0.413
Number of meta-analysis	533	533	533

Notes: $^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$. For all models, our dependent variable is $ESR_{sig.}$, that is, the estimated extent of selective reporting expressed as a share of the significant p-values at the 0.05 level. Standard errors are in parentheses and clustered at the level of meta-studies. The sample is restricted to $ESR_{sig.}$ larger than the 2.5% quantile. 'Experimental' is an indicator variable denoting whether the designs of primary studies included in the meta-analysis are experimental. For the subfield, 'Health, Toxicology and Mutagenesis' is considered as a reference category.

being statistically significant. The regression analysis shows that increasing statistical power and utilizing experimental research designs can help reduce the extent of selective reporting.

With more than half of the significant p-values being falsely reported as statistically significant, the users of empirical environmental science research from scientific peers to policymakers and other stakeholders face substantial uncertainty regarding the reliability of the reported findings. A lack of reliable research results could jeopardize scientific progress and knowledge accumulation, undermine the process of evidence-based policy, and erode trust into science as a whole.

We emphasize that a selectively reported p-value does not necessarily imply that the analyzed effect is absent. In many cases, the hypothesized effect may exist, but the sample size is too small to reliably detect it. The process of selective reporting then results in exaggerated effect sizes, as large estimated effect sizes can ensure statistical significance even if the sample size is small. Specifically, in economics it was found that nearly 80% of effect sizes are exaggerated at least twofold (Ioannidis et al. 2017). Such an exaggeration of effect sizes poses a fundamental challenge for evidence-based policy and decision making: Even though the analyzed effect might exist, it is likely to be substantially smaller than can be inferred from the published literature or may even be practically irrelevant (McCloskey and Ziliak 1996).

Increasing sample sizes and routinely incorporating power considerations into empirical analyses is likely to greatly reduce the extent of selective reporting (Newcombe 1987; Lamberink et al. 2018). However, not all research questions allow the researcher to control sample size, but our data suggest a large potential for improvement. However, experimental research designs, which usually allow the researcher to control sample size, have less power in our sample than research that is not based on randomization ($r = -0.24$). It is essential to improve power where this can be achieved relatively easily and there is a need for highly powered field studies (Yang et al. 2022). As our regression analysis also indicated, adopting experimental research designs may help minimize the extent of selective reporting as opposed to observational research designs (Brodeur et al. 2020; Bruns et al. 2019; Christie et al. 2020). However, it is important to emphasize that neither more power nor improved research designs can reduce selective reporting entirely. Incentives remain to search and select for results that can be published, including large effect sizes, sensational or theory-confirming results, or simply statistically significant results in the absence of a true effect.

In order to further reduce the risk of selective reporting and to generally improve the reliability of empirical research in environmental sciences, all parties involved in the research system need to adhere to transparency and open science as much as possible (Parker et al. 2016; Nosek et al. 2015). In particular, researchers can be transparent by committing to standard reporting guidelines, pre-registering their study protocol(Parker et al. 2019; Soderberg et al. 2021; Lindsley et al. 2022), making raw data and analytical codes accessible to the public (Tedersoo et al. 2021), or blindly collecting and analyzing data (MacCoun and Perlmutter 2015; Holman et al. 2015).

Our analysis again suggests that there is much room for improvement regarding open science. Among the meta-analyses that met our inclusion criteria (701), only 15.5% shared their underlying data online (either in the supplementary information or through data repository platforms). We also contacted the authors of 506 articles for data. Of the 38.6% who responded, only 28.3% provided usable information. Many authors were reluctant to share data for a variety of reasons (e.g., requesting co-authorship as a pre-requisite to sharing the data of published articles).

Journals can also play an important role in promoting the culture of open science and transparency (Hansford et al. 2022). This can be done in a range of ways, such as policies to accept negative or null findings (Blanco-Perez and Brodeur 2020), enforcing data sharing (Askarov et al. 2022), publishing studies based on their scientific merits regardless of whether or not they provide statistically significant findings, encouraging replication studies, and disclosing conflicts of interest (Nosek et al. 2015). Funders can help by insisting on principles of open science and academic institutions can further improve current research practices and reduce selective reporting by providing statistical literacy training, especially for junior researchers (Touchon and McCoy 2016). Communities can establish improved research practices, such as the Open Science Framework (OSF) and the Society for Open, Reliable, and Transparent Ecology and Evolutionary biology (SORTEE) (O'Dea et al. 2021). Ultimately, incentives and research practices need to evolve to align 'getting it published' with 'getting it right' (Nosek et al. 2012).

Most of the differences between the subfields of environmental sciences can be attributed to methodological differences in the use of experimental research designs and statistical power. The subfield 'Health, Toxicology and Mutagenesis' performed best among those subfields with 20 or more meta-analyses with regard to both selective reporting and statistical power. Health-related research started early to reflect on research practices and to conduct research-on-research (Ioannidis 2005). The recent uptake of meta-research in ecology and evolutionary biology (Parker et al. 2019; Yang et al. 2022; Spake et al. 2022; Nakagawa and Parker 2015), together with the decrease of selective reporting over time as suggested by our regression analysis, provides a positive prospect for the reliability of future research in the environmental sciences.

Methods

Literature search

We performed a literature search in Scopus via Hasselt University's institutional access to find meta-analyses in the environmental sciences. We opted to focus on Scopus because we use the research field taxonomy developed by SCImago Journal and Country Rank (SJR) to classify environmental science meta-analyses into subfields, which we discuss further in the following section. The Web of Science (WoS) does not offer a classification scheme that categorizes the environmental sciences into subfields. The search string discussed below identifies more meta-analyses in the WoS compared to Scopus (570 additional hits), which is mainly driven by the *Journal of Clinical Epidemiology* being classified as an environmental science journal in WoS but not in Scopus (440 hits). Generally, Scopus indexes more journals than (Mongeon and Paul-Hus 2016).

We limited the search to meta-analyses published between 2010 and 2020 as the culture and requirements of data-sharing, which have gained momentum in recent years, increase the probability of the primary studies data being publicly available. We used the search string TITLE-ABS-KEY ("meta analy" OR "meta-analy" OR "metaanaly" OR "meta reg" OR "meta-reg" OR "metareg") AND SUBJAREA (envi) to identify relevant meta-analyses. Excluding publications written in a language other than English and considering only article, review, and early access document types, the search query results in 4867 potential records on July 21, 2020.

We also checked for meta-analyses that are hidden in systematic reviews without the word 'meta-analysis' being mentioned in the title, abstract, or keywords. We used the search string TITLE-ABS-KEY ("systematic review" AND NOT ("meta analy" OR "meta-analy" OR "metaanaly*")) AND SUBJAREA (envi), which yields 2814 hits. A random sample of 280 ($\sim 10\%$) papers were screened for their eligibility against a priori defined inclusion and exclusion criteria. We find only eight (2.9%) documents that are deemed eligible for further scrutiny. The vast majority of articles in this sample show that systematic reviews that do not mention 'meta-analysis' in the title, abstract, or keywords indeed do not conduct a meta-analysis. Therefore, we decided to focus on articles that mention some variant of 'meta-analysis' in either title, abstract, or keywords.

Classification of meta-analyses and sampling strategy

We assign each meta-analysis to one of the twelve subfields of environmental sciences that are offered by SJR: 'Ecology', 'Health, Toxicology and Mutagenesis', 'Nature and Landscape Conservation', 'Environmental Chemistry', 'Environmental Engineering', 'Management, Monitoring, Policy and Law', 'Water Science and Technology', 'Waste Management and Disposal', 'Pollution', 'Global and Planetary Change', 'Ecological Modelling', and 'Environmental Science (miscellaneous)'. We use these subfields for two reasons. First, categorization into subfields permits us to draw a stratified random sample to ensure a comprehensive picture of environmental sciences. Second, subfields may explain variations in the extent of selective reporting and the levels of statistical power due to different research practices, methods, and topics.

In the SJR classification, each journal is assigned to at least one subfield. Many journals are assigned to multiple subfields, including subfields outside environmental sciences. For example, a meta-analysis may be published in a journal that is classified as 'Ecology' and 'Transportation'. This meta-analysis is identified by our search string but may actually be a 'Transportation' meta-analysis. Therefore, we determine the subfield of each meta-analysis by using the subfields of the journals listed in the references of this meta-analysis (Milojevic 2020). The implicit assumption is that meta-analyses predominantly cite articles from the subfield that they synthesize. The SJR portal provides a list of 1659 journals, book series, and conference proceedings that are classified into at least one subfield of environmental sciences (www.scimagojr.com). The pseudocode to determine whether an article is an environmental sciences article and to which subfield of environmental sciences it belongs is as follows:

1. All cited references of the article are extracted from Scopus and exported to Excel.
2. The journal, book series, and conference proceeding names related to the field of environmental sciences are identified from the reference list using text mining.
3. The share of journals, book series, and conference proceedings that are classified in at least one environmental sciences subfield is calculated for the reference list of each article. If this share is larger than 25% (median value), the article is considered to be an environmental science article.
4. If the article is classified as an environmental science article, its specific subfield is determined by using the subfield that occurs most frequently in the references (using weights). Specifically, if a journal in the reference list is assigned to four distinct subfields, then each subfield is assigned a weight of 0.25 while for a journal with two subfields each subfield is assigned a weight of 0.5, and for a journal with only one subfield the subfield is assigned a weight of 1.
5. In case a tie occurs between the most frequent subfield and the second most frequent subfield, the original SJR subfield(s) (those of the journal in which the article is published) are added to the count to break ties in line with the work of Milojevic (2020) and WoS recommendations.

As a result, 1706 records were identified as articles from environmental sciences. We followed the stratified sampling strategy employed in Fanelli et al. (2017) and randomly selected 1172 articles (see Figure 3). The full-texts of all selected articles were downloaded and reviewed thoroughly for eligibility, using the inclusion and exclusion criteria outlined in the next section. We randomized the order in which the articles were reviewed to reduce the risk of biases during document screening and data extraction.

Inclusion and exclusion criteria

For a meta-analysis to be included, the following criteria need to be met: (a) raw data for primary studies are available (directly online or after sending a request to the first and last author with a reminder after three weeks, see the following section), or information provided is sufficient to calculate the effect size of interest along with its standard error, (b) published between 2010 and July 21, 2020, (c) the meta-analysis contains at least five independent primary studies, and (d) only article, review, or early access document type. An article is excluded if it is: (a) written in a language other than English, (b) a methodological and/or simulation-based meta-analysis, (c) a reanalysis, comment, letter, editorial, guideline, lecture, book, book chapter, or reply to previous study, (d) a meta-analysis of meta-analyses, (f) electronically inaccessible (e.g., irretrievable due to dead external links or subscription problem), or (g) it does not contain a formal quantitative synthesis (e.g., narrative review, systematic review, vote counting, review protocol, 'unconventional' meta-analysis).

These inclusion and exclusion criteria led to the inclusion of 247 meta-analyses in the final sample. We reclassified 40 meta-analyses assigned to the Environmental Science (miscellaneous) category into other existing categories. Only one meta-analysis was classified as Pollution. This meta-analysis deals with the transmission and accumulation of trace metals in marine food webs and we reclassified it as Environmental Chemistry.

As many meta-studies report more than one meta-analysis, at least 10 meta-analyses are available for each of the seven subfields where the sampling strategy provided at least one meta-study. For the analyses, we obtain standard errors and confidence intervals by clustering at the level of meta-studies to account for dependence of meta-analyses published in the same meta-study (Abadie et al. 2017).

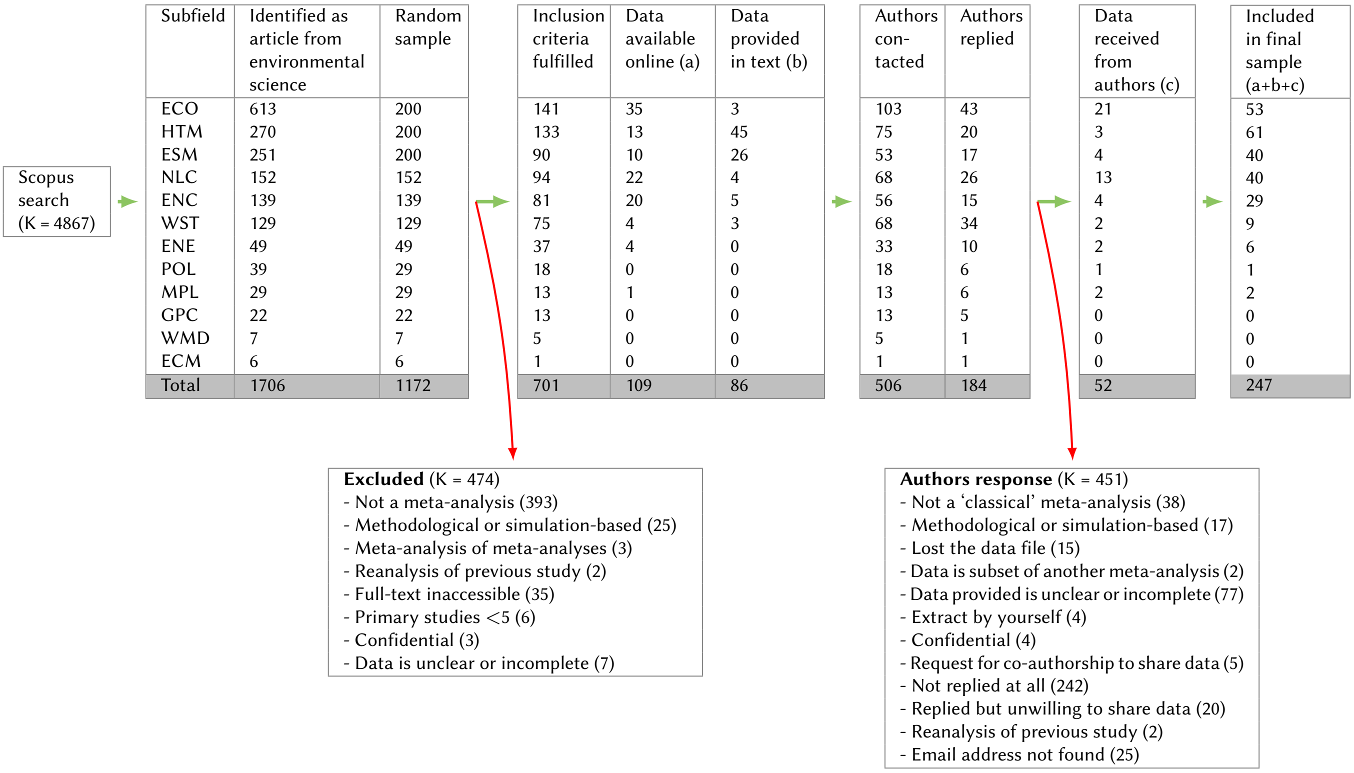
It is important to emphasize that the classification into subfields is not unambiguous. Researchers may easily disagree whether a meta-analysis belongs to one subfield or the other, or to both. For example, the meta-analysis 'Effects of land management on the abundance and richness of spiders (Araneae): A meta-analysis' may easily be classified as 'Ecology' or 'Nature and Landscape Conservation'. We apply an automated classification that helps to ensure a transparent and objective classification into subfields. Moreover, a sample of 134 meta-studies (54.25%) was checked by four co-authors who are experts in the respective subfield to determine whether the automated classification resulted in gross misclassifications or whether the resulting classification into subfields can be considered reasonable. For 8.96% of the meta-analyses checked, a different subfield was considered to be more appropriate, but disagreement on this can also not be excluded.

Extraction of data

From the meta-analyses, we extracted effect sizes and standard errors of the primary studies (or alternative information to calculate these) using the following strategy:

1. We searched the journal's webpage for online supplements.
2. If the journal webpage does not provide the effect sizes and standard errors of the primary studies, we searched for links in the meta-analysis to online data repository platforms (e.g., Open Science Framework (OSF), Zenodo, figshare, Harvard Dataverse, Dryad, Ecological Archives) by using the search strings "http" and "www" in the meta-analysis.
3. If we were unable to find a link, the data is not given in the meta-analysis, or the data provided is unclear to calculate effect sizes and corresponding standard errors, we contacted the authors to request raw data or clarification. Using an email with standard text but customized by specifying the authors' name, the meta-analysis title, and the journal name, we contacted two authors (the first and last in the order they are listed in the meta-analysis). We sent a reminder email to those who did not reply after three weeks. If none of the contacted authors replied the meta-analysis was excluded. Contacting authors resulted in additional primary data for 52 meta-analyses. Details with regard to the authors response disaggregated by subfields are given in Figure 3.

For each included meta-analysis, we coded the title of the meta-analysis, publication year, journal name, number of independent primary studies, number of effect sizes, subfield, whether the authors formally adhered



Scopus search (K = 4867)

Subfield	Identified as article from environmental science	Random sample	Inclusion criteria fulfilled	Data available online (a)	Data provided in text (b)	Authors con- tacted	Authors replied	Data received from authors (c)	Included in final sample (a+b+c)
ECO	613	200	141	35	3	103	43	21	53
HTM	270	200	133	13	45	75	20	3	61
ESM	251	200	90	10	26	53	17	4	40
NLC	152	152	94	22	4	68	26	13	40
ENC	139	139	81	20	5	56	15	4	29
WST	129	129	75	4	3	68	34	2	9
ENE	49	49	37	4	0	33	10	2	6
POL	39	29	18	0	0	18	6	1	1
MPL	29	29	13	1	0	13	6	2	2
GPC	22	22	13	0	0	13	5	0	0
WMD	7	7	5	0	0	5	1	0	0
ECM	6	6	1	0	0	1	1	0	0
Total	1706	1172	701	109	86	506	184	52	247

Excluded (K = 474)
- Not a meta-analysis (393)
- Methodological or simulation-based (25)
- Meta-analysis of meta-analyses (3)
- Reanalysis of previous study (2)
- Full-text inaccessible (35)
- Primary studies <5 (6)
- Confidential (3)
- Data is unclear or incomplete (7)

Authors response (K = 451)
- Not a ‘classical’ meta-analysis (38)
- Methodological or simulation-based (17)
- Lost the data file (15)
- Data is subset of another meta-analysis (2)
- Data provided is unclear or incomplete (77)
- Extract by yourself (4)
- Confidential (4)
- Request for co-authorship to share data (5)
- Not replied at all (242)
- Replied but unwilling to share data (20)
- Reanalysis of previous study (2)
- Email address not found (25)

Figure 3: Flowchart of the study selection process. Exclusion of selected meta-analyses is carried out after a review of the full text. ECO = Ecology, HTM = Health, Toxicology and Mutagenesis, ESM = Environmental Science (miscellaneous), NLC = Nature and Landscape Conservation, ENC = Environmental Chemistry, WST = Water Science and Technology, ENE = Environmental Engineering, POL = Pollution, MPL = Management, Monitoring, Policy and Law, GPC = Global and Planetary Change, WMD = Waste Management and Disposal, ECM = Ecological Modelling.

to any review guidelines or checklist (e.g., PRISMA, ROSES, MOOSE), and whether the meta-analysis is pre-registered.

The application of the inclusion and exclusion criteria and the data extraction was conducted by the first author and supervised by the last author. Disagreements were resolved by discussion until a consensus was reached.

Analytical approach

Power computation

In order to calculate the power for each statistical test in each meta-analysis, the respective genuine effect needs to be estimated. We use the weighted average of the adequately powered (WAAP) studies whenever there are at least two sufficiently powered primary estimates in a given meta-analysis (Ioannidis et al. 2017).[1]

If there are fewer than two adequately powered estimates, the WAAP estimator uses the precision effect test and the precision effect estimate with standard error (PET-PEESE) (Stanley and Doucouliagos 2014). We compute power assuming all tests originate from a two-sided z-test using a 5% significance level:

$$\text{Power} = 1 - [\Phi(z_{\alpha/2} - \lambda)] - \Phi(-z_{\alpha/2} - \lambda) \tag{1}$$

where $\Phi()$ is the cumulative standard normal distribution, $\lambda = \hat{\beta}^{*}/se$ is the non-centrality parameter with $\hat{\beta}^{*}$ being the estimated genuine effect and se the standard error of the respective primary estimate and $z_{\alpha/2}$ is

[1] We excluded 7957 observations with an absolute z-value larger than 20 to reduce the risk that estimated meta-averages are inflated by outliers. Sensitivity analysis also explores thresholds of 50 and 100. Please see Figure S1 in the Online Appendix.

the critical z-value for $\alpha = 0.05$. We then provide information on the median power for each meta-analysis and the share of adequately powered studies per meta-analysis. Adequate power is defined as a power of 0.8 (Ioannidis et al. 2017).

Construction of counterfactual z- and p-values

Reported p-values are frequently subject to selective reporting (e.g., Bruns et al. 2019; Bruns et al. 2022; Brodeur et al. 2020). To estimate the extent of selective reporting, we use the recently proposed approach by Bruns et al. (2022) that compares the empirical distribution of published p-values with the counterfactual distribution generated in the absence of any biases. Comparing these two distributions allows us to infer where p-values are missing and where p-values are overrepresented. We rely on three assumptions to construct the counterfactual distribution of p-values:

(i) There is one genuine effect per meta-analysis.

(ii) Genuine effects can be approximated by meta-analytic estimators.

(iii) Standard errors are not affected by selective reporting.

We conduct comprehensive robustness analyses for each of these three assumptions. These robustness analyses are discussed at the end of this section.

Following Bruns et al. (2022), let β_j and $\hat{\beta}_{ji}$ be the genuine effect for meta-analysis j and the corresponding estimate obtained from primary study i, respectively. If all studies estimate the respective genuine effect unbiasedly, then $E(\hat{\beta}_{ji}) = \beta_j$ holds. Consequently, $\hat{\beta}_{ji} \sim \mathcal{N}(\beta_j, se_{ji})$, where se_{ji} is the true standard error of estimate i in meta-analysis j. We can then simulate counterfactual z-values (z^{cf}) as

$$z^{cf} \sim \mathcal{N}\left(\frac{\beta_j}{se_{ji}}, 1\right). \qquad (2)$$

In a given interval $[a, b]$, the expected frequency of counterfactual z-values is given by

$$E[r^{cf}_{a,b}] = \sum_{j}^{K} \sum_{i}^{S_j} \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(x - \left(\frac{\beta_j}{se_{ji}}\right)\right)^2}{2}\right) dx \qquad (3)$$

where K with $j = 1, \cdots, K$ is the number of meta-analyses and S_j with $i = 1, \cdots, S_j$ is the number of primary estimates in meta-analysis j. Subsequently, assuming all tests originate from a two-sided z-test, the expected frequency of counterfactual p-values in the interval $[g, h]$ is given by

$$E[f^{cf}_{g,h}] = E[r^{cf}_{-b,-a}] + E[r^{cf}_{a,b}] \qquad (4)$$

where $a = Q(1 - h/2), b = Q(1 - g/2)$, and Q is the quantile function of the standard normal distribution. We can then estimate $E[f^{cf}_{g,h}]$ by plugging in $\hat{\beta}^*_j$ for β_j and $\hat{se}_{ji}$ for se_{ji} in Eq. (3).

The resulting counterfactual distribution can be compared to the distribution of observed p-values using relative differences:

$$D_{g,h} = \frac{f^f_{g,h} - E[f^{cf}_{g,h}]}{f^f_{0,1}} \qquad (5)$$

where $f^f_{g,h}$ is the frequency of reported p-values in the interval $[g, h]$ and $f^f_{0,1}$ is the total number of published p-values. This relative difference allows us to assess the abundance or lack of p-values in any interval over the entire support of p-values.

We assess the extent of selective reporting by comparing the frequency of published and statistically significant p-values to the expected frequency of significant p-values. The extent of selective reporting, ESR, is estimated by

$$ESR_s = f_s^f - E[f_s^{cf}] \tag{6}$$

where the subscript s denotes a pre-specified level of significance. We then express the extent of selective reporting as a share of all p-values and as a share of the significant p-values.

To explore robustness with regard to the assumption of one genuine effect per meta-analysis, we randomly split the primary estimates in each meta-analysis into two groups and then estimate the meta-averages separately for each group by means of WLS. We also conduct this robustness check by randomly splitting into three groups. To examine the robustness with regard to the estimation of the meta-average, we use four alternative methods: WLS, PET-PEESE, a bias-corrected stem-based method proposed by Furukawa (2019), and half the meta-average obtained by WAAP. The motivation for using half the meta-average is that whenever there is evidence of selective reporting, estimated meta-averages are known to overestimate the true effect (Bruns et al. 2022; Yang et al. 2022). Overestimating the meta-average results in an underestimation of the extent of selective reporting. With regard to the assumption that the standard errors are not subject to selective reporting, we conduct the analysis again but by multiplying standard errors by 1.5 and 2. Finally, we also explore robustness with regard to excluding observations with $|z| > 50$ and $|z| > 100$.

Exploratory regression analysis

We also explore potential determinants of selective reporting. We fit regression models to explore whether characteristics of the primary studies and meta-analyses or differences between the various subfields explain variation in the extent of selective reporting expressed as a share of significant p-values. The following model is considered (see Table 5 for the description of variables):

$$\begin{aligned} \text{ESR}_j = {} & \beta_0 + \beta_1 \text{SAPE}_j + \beta_2 \text{RDES}_j + \beta_3 \text{GUID}_j + \beta_4 \text{PRER}_j + \boldsymbol{\beta}_5 \textbf{SUBF}_j \\ & + \beta_6 \log(\text{NPRS}_j) + \beta_7 \log(\text{NTPV}_j) + \beta_8 \text{YOP}_j \end{aligned} \tag{7}$$

where $j = 1, \cdots, K$ indexes meta-analyses. The share of adequately powered estimates (SAPE) and the design of primary studies (RDES) measure characteristics of the primary studies with regard to the sample sizes and corresponding power as well as whether experimental or observational research designs were employed. Whether standard guidelines were followed (GUID) and whether the meta-analysis was pre-registered (PRER) measure quality characteristics at the level of the meta-analysis. A vector of subfields dummies captures whether differences between subfields can explain variation in the extent of selective reporting, the number of primary studies (NPRS), the total number of p-values (NTPV), and the publication year (YOP) are included as control variables.

Further, we fitted Eq. (7) by including the interactions between research design (RDES) and the share of adequately powered estimates (SAPE). The reason is that RDES and SAPE are negatively correlated in the sample and we expect RDES and SAPE to have a negative effect on ESR. We cluster standard errors at the level of meta-studies to account for potential dependence of meta-analysis published in the same meta-study.

For the regression analysis, we excluded from the distribution of $ESR_{sig.}$ a few outliers. Specifically, we excluded meta-analysis below the 2.5% quantile of the $ESR_{sig.}$ distribution. The reason is that meta-averages can be easily overestimated in the presence of selective reporting and this may then result in negative estimates of ESR_{sig}. In other words, if the meta-average is strongly overestimated, then even statistical tests reported as being non-significant are expected to be statistically significant, which can turn $ESR_{sig.}$ to become negative. The 2.5% quantile of the $ESR_{sig.}$ distribution corresponds to $ESR_{sig.} = -0.61$ and 14 meta-analyses are excluded. As robustness checks, we run the analysis with all meta-analyses, using half the meta-average as a proxy of the genuine effect and by excluding outliers using the 5% quantile of the $ESR_{sig.}$ distribution corresponding to 28 meta-analyses and $ESR_{sig.} = -0.35$. If all meta-analyses are used, the effect of the interaction term and publication year vanishes. The sign and significance of the variables remain largely the same for the other robustness checks (see Tables S8–S10 in the Online Appendix).

Table 5: Description of study variables.

Variable	Description
ESR	Extent of selective reporting estimated using Eq. (6) with $s = 0.05$
SAPE	Share of adequately powered estimates in meta-analysis j
RDES	Design of studies included in meta-analysis j (experimental *vs.* observational)
NPRS	Number of independent primary studies included in meta-analysis j
SUBF	Subfield of meta-analysis j (see Figure 3 for subfields list and corresponding labels)
GUID	Whether meta-analysis j adhered to standard guidelines such as PRISMA and ROSES (yes, no)
PRER	Whether meta-analysis j is pre-registered and mentions this in the main report or supplementary material (yes, no)
NTPV	Total number of p-values in meta-analysis j
YOP	The year meta-analysis j is published

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge. 2017. *When should you adjust standard errors for clustering?* Technical report. National Bureau of Economic Research.

Albarqouni, Loai N, José A López-López, and Julian PT Higgins. 2017. "Indirect evidence of reporting biases was found in a survey of medical research studies." *Journal of Clinical Epidemiology* 83:57–64.

Askarov, Zohid, Anthony Doucouliagos, Hristos Doucouliagos, and T D Stanley. 2022. "The Significance of Data-Sharing Policy." *Journal of the European Economic Association,* ISSN: 1542-4766. https://doi.org/10.1093/jeea/jvac053.

Barnes, JC, Michael F TenEyck, Travis C Pratt, and Francis T Cullen. 2020. "How Power Ful Is the Evidence in Criminology? On whether We Should Fear a Coming Crisis of Confidence." *Justice Quarterly* 37 (3): 383–409.

Bastardi, Anthony, Eric Luis Uhlmann, and Lee Ross. 2011. "Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence." *Psychological Science* 22 (6): 731.

Baum, Joel, and Philip Bromiley. 2019. "P-hacking in Top-tier Management Journals." In *Academy of Management Proceedings,* 2019:10810. 1. Academy of Management Briarcliff Manor, NY 10510.

Blanco-Perez, Cristina, and Abel Brodeur. 2020. "Publication bias and editorial statement on negative findings." *The Economic Journal* 130 (629): 1226–1247.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, et al. 2020. "Variability in the analysis of a single neuroimaging dataset by many teams." *Nature* 582 (7810): 84–88.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung HV Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, Henrik K Andersen, Daniel Auer, Flavio Azevedo, et al. 2022. "Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty." *Proceedings of the National Academy of Sciences* 119 (44): e2203150119.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods matter: P-hacking and publication bias in causal analysis in economics." *American Economic Review* 110 (11): 3634–60.

Bruns, Stephan B, Igor Asanov, Rasmus Bode, Melanie Dunger, Christoph Funk, Sherif M Hassan, Julia Hauschildt, Dominik Heinisch, Karol Kempa, Johannes Konig, et al. 2019. "Reporting errors and biases in published empirical findings: Evidence from innovation research." *Research Policy* 48 (9): 103796.

Bruns, Stephan B, Teshome K Deressa, TD Stanley, Hristos Doucouliagos, and John PA Ioannidis. 2022. "Estimating the extent of inflated significance in economics." *MetaArXiv,* https://doi.org/10.31222/osf.io/h29xn.

Bruns, Stephan B, and Martin Kalthaus. 2020. "Flexibility in the selection of patent counts: Implications for p-hacking and evidence-based policymaking." *Research Policy* 49 (1): 103877.

Button, Katherine S, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. "Power failure: why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience* 14 (5): 365–376.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jurgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. 2016. "Evaluating replicability of laboratory experiments in economics." *Science* 351 (6280): 1433–1436.

Chang, Andrew C, and Phillip Li. 2022. "Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not"." *Critical Finance Review* 11 (1): 185–206.

Chen, Yuyu, Ginger Zhe Jin, Naresh Kumar, and Guang Shi. 2012. "Gaming in air pollution data? Lessons from China." *The BE Journal of Economic Analysis & Policy* 13 (3).

Christie, Alec P, David Abecasis, Mehdi Adjeroud, Juan C Alonso, Tatsuya Amano, Alvaro Anton, Barry P Baldigo, Rafael Barrientos, Jake E Bicknell, Deborah A Buhl, et al. 2020. "Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences." *Nature Communications* 11 (1): 1–11.

Cleasby, Ian R, Barbara J Morrissey, Mark Bolton, Ellie Owen, Linda Wilson, Saskia Wischnewski, and Shinichi Nakagawa. 2021. "What is our power to detect device effects in animal tracking studies?" *Methods in Ecology and Evolution* 12 (7): 1174–1185.

Dumas-Mallet, Estelle, Katherine S Button, Thomas Boraud, Francois Gonon, and Marcus R Munafò. 2017. "Low statistical power in biomedical science: a review of three human research domains." *Royal Society Open Science* 4 (2): 160254.

Fanelli, Daniele, Rodrigo Costas, and John PA Ioannidis. 2017. "Meta-assessment of bias in science." *Proceedings of the National Academy of Sciences* 114 (14): 3714–3719.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345 (6203): 1502–1505.

Furman, Jeffrey L, Kyle Jensen, and Fiona Murray. 2012. "Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine." *Research Policy* 41 (2): 276–290.

Furukawa, Chishio. 2019. "Publication bias under aggregation frictions: Theory, evidence, and a new correction method." *Evidence, and a New Correction Method (March 29, 2019).*

Gerber, Alan, and Neil Malhotra. 2008b. "Do statistical reporting standards affect what is published? Publication bias in two leading political science journals." *Quarterly Journal of Political Science* 3 (3): 313–326.

———. 2008a. "Publication bias in empirical sociological research: Do arbitrary significance levels distort published results?" *Sociological Methods & Research* 37 (1): 3–30.

Haddaway, Neal R, Biljana Macura, Paul Whaley, and Andrew S Pullin. 2018. "ROSES RepOrting standards for Systematic Evidence Syntheses: pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps." *Environmental Evidence* 7 (1): 1–8.

Hansford, Harrison J, Aidan G Cashin, Michael A Wewege, Michael C Ferraro, James H McAuley, and Matthew D Jones. 2022. "Open and transparent sports science research: the role of journals to move the field forward." *Knee Surgery, Sports Traumatology, Arthroscopy,* 1–3.

Harrison, Jeffrey S, George Christopher Banks, Jeffrey M Pollack, Ernest H O'Boyle, and Jeremy Short. 2017. "Publication bias in strategic management research." *Journal of Management* 43 (2): 400–425.

Holman, Luke, Megan L Head, Robert Lanfear, and Michael D Jennions. 2015. "Evidence of experimental bias in the life sciences: why we need blind data recording." *PLoS Biology* 13 (7): e1002190.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, et al. 2021. "The influence of hidden researcher decisions in applied microeconomics." *Economic Inquiry* 59 (3): 944–960.

Ioannidis, John. 2005. "Why most published research findings are false." *PLoS Medicine* 2 (8): e124.

Ioannidis, John, TD Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal* 127 (605): F236–F265.

Lamberink, Herm J, Willem M Otte, Michel RT Sinke, Daniel Lakens, Paul P Glasziou, Joeri K Tijdink, and Christiaan H Vinkers. 2018. "Statistical power of clinical trials increased while effect size remained stable: an empirical analysis of 136,212 clinical trials between 1975 and 2014." *Journal of Clinical Epidemiology* 102:123–128.

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

Lindsley, Kristina, Nicole Fusco, Tianjing Li, Rob Scholten, and Lotty Hooft. 2022. "Clinical trial registration was associated with lower risk of bias compared with non-registered trials among trials included in systematic reviews." *Journal of Clinical Epidemiology* 145:164–173.

MacCoun, Robert, and Saul Perlmutter. 2015. "Blind analysis: Hide results to seek the truth." *Nature News* 526 (7572): 187.

McCloskey, Deirdre N, and Stephen T Ziliak. 1996. “The standard error of regressions.” *Journal of Economic Literature* 34 (1): 97–114.

Milojevic, Staša. 2020. “Practical method to reclassify Web of Science articles into unique subject categories and broad disciplines.” *Quantitative Science Studies* 1 (1): 183–206.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and PRISMA Group*. 2009. “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.” *Annals of Internal Medicine* 151 (4): 264–269.

Mongeon, Philippe, and Adèle Paul-Hus. 2016. “The journal coverage of Web of Science and Scopus: a comparative analysis.” *Scientometrics* 106 (1): 213–228.

Nakagawa, Shinichi, and Timothy H Parker. 2015. “Replicating research in ecology and evolution: feasibility, incentives, and the cost-benefit conundrum.” *BMC Biology* 13 (1): 1–6.

Newcombe, Robert G. 1987. “Towards a reduction in publication bias.” *British Medical Journal (Clinical Research Edition)* 295 (6599): 656–659.

Nickerson, Raymond S. 1998. “Confirmation bias: A ubiquitous phenomenon in many guises.” *Review of General Psychology* 2 (2): 175–220.

Nosek, Brian A, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. 2015. “Promoting an open research culture.” *Science* 348 (6242): 1422–1425.

Nosek, Brian A, Jeffrey R Spies, and Matt Motyl. 2012. “Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability.” *Perspectives on Psychological Science* 7 (6): 615–631.

Nuijten, Michèle B, Marcel ALM van Assen, Hilde EM Augusteijn, Elise AV Crompvoets, and Jelte M Wicherts. 2020. “Effect sizes, power, and biases in intelligence research: a meta-meta-analysis.” *Journal of Intelligence* 8 (4): 36.

O’Dea, Rose E, Timothy H Parker, Yung En Chee, Antica Culina, Szymon M Drobniak, David H Duncan, Fiona Fidler, Elliot Gould, Malika Ihle, Clint D Kelly, et al. 2021. “Towards open, reliable, and transparent ecology and evolutionary biology.” *BMC Biology* 19 (1): 1–5.

Open Science Collaboration. 2015. “Estimating the reproducibility of psychological science.” *Science* 349 (6251): 253–267. ISSN: 0036-8075.

Parker, Timothy, Wolfgang Forstmeier, Julia Koricheva, Fiona Fidler, Jarrod D Hadfield, Yung En Chee, Clint D Kelly, Jessica Gurevitch, and Shinichi Nakagawa. 2016. “Transparency in ecology and evolution: real problems, real solutions.” *Trends in Ecology & Evolution* 31 (9): 711–719.

Parker, Timothy, Hannah Fraser, and Shinichi Nakagawa. 2019. “Making conservation science more reliable with preregistration and registered reports.” *Conservation Biology* 33 (4): 747–750.

Pietschnig, Jakob, Magdalena Siegel, Junia Sophia Nur Eder, and Georg Gittler. 2019. “Effect declines are systematic, strong, and ubiquitous: A meta-meta-analysis of the decline effect in intelligence research.” *Frontiers in Psychology* 10:2874.

Rosenthal, Robert. 1979. “The file drawer problem and tolerance for null results.” *Psychological Bulletin* 86 (3): 638.

Rost, Katja, and Thomas Ehrmann. 2017. “Reporting biases in empirical management research: The example of win-win corporate social responsibility.” *Business & Society* 56 (6): 840–888.

Silberzahn, Raphael, Eric L Uhlmann, Daniel P Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Stepan Bahnik, Feng Bai, Colin Bannard, Evelina Bonnier, et al. 2018. “Many analysts, one data set: Making transparent how variations in analytic choices affect results.” *Advances in Methods and Practices in Psychological Science* 1 (3): 337–356.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. “False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant.” *Psychological Science* 22 (11): 1359–1366.

Smith, Daniel R, Ian CW Hardy, and Martin P Gammell. 2011. "Power rangers: no improvement in the statistical power of analyses published in Animal Behaviour." *Animal Behaviour* 1 (81): 347–352.

Soderberg, Courtney K, Timothy M Errington, Sarah R Schiavone, Julia Bottesini, Felix Singleton Thorn, Simine Vazire, Kevin M Esterling, and Brian A Nosek. 2021. "Initial evidence of research quality of registered reports compared with the standard publishing model." *Nature Human Behaviour* 5 (8): 990–997.

Spake, Rebecca, Rose E O'Dea, Shinichi Nakagawa, C Patrick Doncaster, Masahiro Ryo, Corey T Callaghan, and James M Bullock. 2022. "Improving quantitative synthesis to achieve generality in ecology." *Nature Ecology & Evolution,* 1–11.

Stanley, Tom D, Evan C Carter, and Hristos Doucouliagos. 2018. "What meta-analyses reveal about the replicability of psychological research." *Psychological Bulletin* 144 (12): 1325.

Stanley, Tom D, and Hristos Doucouliagos. 2014. "Meta-regression approximations to reduce publication selection bias." *Research Synthesis Methods* 5 (1): 60–78.

Stanley, Tom D, Hristos Doucouliagos, and John PA Ioannidis. 2022. "Retrospective median power, false positive meta-analysis and large-scale replication." *Research Synthesis Methods* 13 (1): 88–108.

Tedersoo, Leho, Rainer Küngas, Ester Oras, Kajar Köster, Helen Eenmaa, Äli Leijen, Margus Pedaste, Marju Raju, Anastasiya Astapova, Heli Lukner, et al. 2021. "Data sharing practices and data availability upon request differ across scientific disciplines." *Scientific Data* 8 (1): 1–11.

Touchon, Justin C, and Michael W McCoy. 2016. "The mismatch between current statistical practice and doctoral training in ecology." *Ecosphere* 7 (8): e01394.

Turner, Erick H, Annette M Matthews, Eftihia Linardatos, Robert A Tell, and Robert Rosenthal. 2008. "Selective publication of antidepressant trials and its influence on apparent efficacy." *New England Journal of Medicine* 358 (3): 252–260.

Vivalt, Eva. 2019. "Specification searching and significance inflation across time, methods and disciplines." *Oxford Bulletin of Economics and Statistics* 81 (4): 797–816.

Yang, Yefeng, Helmut Hillebrand, Malgorzata Lagisz, Ian Cleasby, and Shinichi Nakagawa. 2022. "Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology." *Global Change Biology* 28 (3): 969–989.

Zhong, Shifa, Kai Zhang, Majid Bagheri, Joel G Burken, April Gu, Baikun Li, Xingmao Ma, Babetta L Marrone, Zhiyong Jason Ren, Joshua Schrier, et al. 2021. "Machine learning: new ideas and tools in environmental science and engineering." *Environmental Science & Technology* 55 (19): 12741–12754.