More reliable inference from open raw data: a Monte Carlo comparison of raw-data and aggregate-data meta-analysis under publication bias

Danijela Žanko[1], Sunčana Geček[1], Azra Tafro[2], Livia Puljak[3], Antica Culina[1,*]

[1]Laboratory for Informatics and Environmental Modelling, Division for Marine and Environmental Research, Ruđer Bošković Institute, Bijenička Cesta 54, Zagreb 10000, Croatia
[2]Faculty of Forestry and Wood Technology, University of Zagreb, Svetošimunska cesta 23, Zagreb 10000, Croatia
[3]Center for Evidence-Based Medicine, Catholic University of Croatia, Ilica 244, Zagreb 10000, Croatia
*Corresponding author: Antica Culina (aculina@irb.hr)

Keywords: meta-analysis, publication bias, raw data, simulation, effect size estimation, open science

Abstract

While the benefits of open data are often discussed, they are rarely quantified. Here, we provide the first simulation-based quantification of what raw data from all conducted primary studies, including unpublished ones, can add to meta-analysis under publication bias, and we introduce a tool that helps researchers determine when this approach is most beneficial. Classical meta-analysis (CMA) relies on published results and is thus vulnerable to publication bias and p-hacking. We developed a Monte Carlo simulation framework that compares CMA with raw data meta-analysis (RDMA) across varying values of true effect sizes, heterogeneity, and significance rates (the percentage of significant results among published studies). In our simulation we start from a population of conducted primary studies. Some of them get published, subject to p-hacking and publication selection. CMA then samples from these, while RDMA samples raw data from the full population of primary studies, irrespective of whether they are published or not, drawing the same number of studies as CMA. Thus, the two

differ only in which studies they sample, not how many. Using ecology as a case study, we demonstrate that RDMA outperforms CMA in most scenarios. When true effects are small and 95% of published results are significant, RDMA reduces relative mean absolute error in effect size estimates by 54–76%. At a 65% significance rate, reductions reach 32–61%. For medium true effects, reductions are 48–71% and 2–35%, respectively. RDMA maintained reliable confidence interval coverage (93–95%) across all scenarios; CMA fell as low as 21%. Crucially, RDMA's errors reflect natural sampling variation by construction, while CMA's reflect systematic bias that persists regardless of the sample size; in practice, selective availability of raw data would attenuate this contrast. The benefits of RDMA reflect its access to an unbiased, non-p-hacked set of primary studies, and not the recomputation of effect sizes from the raw data of the published literature alone. We provide a decision tool for meta-analysts across disciplines to calculate RDMA's benefits. Our results provide quantitative evidence that access to raw data from all conducted studies, including those that remain unpublished, improves meta-analytic accuracy in the presence of publication bias, strengthening the case not merely for open data but for complete open data, that is, raw data archived for all conducted studies, published and unpublished alike.

1 Introduction

Meta-analysis is the main method for quantitatively synthesising existing evidence and is often used to inform future research, interventions, and policies [1–3]. Importantly, the statistical power of individual primary studies is often inadequate to detect true effects and can thus lead to imprecise effect estimates that limit the detection of meaningful effects [4–8]. Meta-analysis aggregates results from independent studies, increasing statistical power and precision [5]. Critically, this approach assumes that the included studies represent an unbiased sample of all research conducted on a given research question and that the included effects capture the true effects well. However, this assumption is often violated because selective publication and p-hacking can lead to overestimated effects and thus misguided decisions when meta-analysis relies on published results alone. Such selective publication and p-hacking are common across scientific disciplines, for example in economics [9], in observational research [10], and across the sciences more broadly [11–13].

Publication bias [14], the preferential publication of statistically significant results, systematically distorts the published literature [9, 15–18]. Because significant results are more likely to be published, researchers might manipulate their analyses to achieve statistical sig-

nificance, which is called p-hacking [19]. Data across disciplines demonstrate that these issues could be very widespread. For example, 94% of p-values reported in display items from the top multidisciplinary journals are significant [20], 96% of biomedical abstracts and full-text articles reporting p-values contain at least one significant result [21], while in ecology 98.9% of published articles report at least one significant result (computed from [22, 23]). Surveys show that researchers commonly apply p-hacking and other questionable research practices: in psychology, self-admission rates for practices such as selective reporting and optional stopping range from 16–63% [24], with comparable rates confirmed across countries [25, 26] and in ecology and evolution [27], engineering [28], and across Dutch academia more broadly [29]. These findings convincingly indicate that researchers might selectively find and report significant effects, while null findings and non-significant results remain inaccessible, leading to a substantial loss of scientific knowledge [30].

Classical meta-analysis (CMA) commonly synthesises effect sizes calculated from the results reported in published studies, where each study contributes one or more effect sizes. Thus, the biases discussed above can lead to overstated meta-analytical means. For example, simulations by Friese and Frankenbach [31] showed that p-hacking and publication bias interact to severely distort meta-analytic effect size estimates, particularly when true effects are small. Empirically, Yang et al. [15] found that correcting for publication bias across 87 ecological and evolutionary meta-analyses reduced effect sizes by at least 0.12 standard deviations on average, with 33 out of 50 initially significant meta-analytic means becoming non-significant after correction. These biases can also lead to confidence intervals with sub-nominal coverage (i.e., intervals labelled as 95% may contain the true effect less than 95% of the time) [15]. Statistical corrections for publication bias exist, but often perform poorly in the presence of effect size heterogeneity [16], which is widespread in different research fields [32–34]. Consequently, even well-designed CMAs may systematically overestimate true effect sizes when relying solely on published results.

Raw data meta-analysis (RDMA) provides an alternative approach where effect sizes are calculated directly from the original datasets of individual studies rather than from summary statistics reported in published articles. This approach is becoming increasingly viable across the range of disciplines, as open data becomes the norm [35]. Open data is often publicly available in repositories, allowing researchers to recompute the effect size for each study using raw observations. These recalculated study-level effects are then synthesised using standard meta-analytic models. This differs from the individual participant data (IPD) meta-analysis often used in clinical research, where participant-level data from multiple studies are combined into a single

dataset and analysed jointly. In RDMA, raw data are used only to recompute the effect sizes within each study, and the meta-analysis still combines study-level estimates rather than pooling observations across studies [36]. In principle, a raw-data meta-analysis may draw on raw data from published studies only, from unpublished studies only, or from a mixture of both, and these scenarios have very different statistical properties. In this study, we simulate the third and most complete case, RDMA across all conducted studies, published and unpublished alike, corresponding to a field in which raw data are archived irrespective of a study's statistical outcome.

Such raw data syntheses are not merely hypothetically possible. Good examples of accessible raw data are coordinated distributed experiments such as the Nutrient Network, in which a common protocol is applied across many sites and every site contributes to a shared dataset irrespective of its individual result [37]. Mandatory data-archiving policies at journals and funders, together with changes in research culture [38–40], are steadily enlarging the pool of primary studies for which raw data are accessible, making RDMA increasingly feasible. However, open data are still often time-consuming to find [41, 42], interpret, and use [43, 44]. Thus, researchers might want to know when to aim to conduct RDMA (or incorporate raw data in their meta-analysis) rather than CMA. Yet it is still not quantified under what conditions, and by how much, RDMA outperforms CMA. Here, we addressed this gap using a simulation-based approach, where the main conditions under which meta-analysis operates can be set to levels that reflect those present in a given research area. These conditions include the magnitude of the true effect size, heterogeneity between studies, typical sample size, and severity of publication bias. In our simulation, CMA is represented by a conventional random-effects meta-analysis of published effect sizes, and not adjusted for publication bias, since this remains the dominant practice in ecology [15]. We aim to quantify the value of access to an unbiased set of conducted studies and their non-p-hacked results, rather than to benchmark RDMA against bias-correction methods. To parameterise the simulation and understand the potential benefits of RDMA, we used ecology as a case example. However, the framework is designed so that researchers can evaluate the performance of RDMA versus CMA in their own domain by modifying the input parameters (effect size magnitude, heterogeneity level, sample size, and significance rate). We provide worked examples using parameters from ecology, complete documentation, and annotated scripts accessible in the online repository [45].

2 Methodology

The simulation design, the simulation parameter values, the metrics to measure the benefits of RDMA, and the planned comparisons between RDMA and CMA performance were preregistered on the Open Science Framework [46]. Deviations are explained in Supplementary Material S1. All code and data are openly available [45].

2.1 The overarching approach

Our framework uses Monte Carlo simulation to compare RDMA and CMA across different scenarios of true effect size magnitude and heterogeneity levels, and under different levels of presence of significant results in the published literature. Each scenario is evaluated over many independent stochastic replicates (10,000 in the main analysis), and performance is assessed from the distribution of estimates across these replicates, which is what makes the procedure a Monte Carlo experiment. For each scenario, the algorithm generates a population of N conducted studies, representing all research on a given research question before any selection for publication occurs. Each study is represented by an effect size (Hedges' g) drawn from a distribution with mean θ (the true effect size) and variance τ^2 (the heterogeneity between studies). We chose Hedges' g as the effect size measure because the standardised mean difference is one of the two most common effect size measures in meta-analysis across disciplines [2, 3, 47, 48]. Under a selective publication process, only a subset of these studies (i.e., of effect sizes) becomes available as the published literature, and we design this process so that it generates target rates of statistical significance among the published studies. This published subset is what CMA samples from. Under a complete data-archiving scenario, by contrast, the raw data (and thus the effect sizes) of all conducted studies are potentially available, and RDMA samples from this full, unselected population. The two methods therefore sample the same number of studies but from different pools: CMA from the published subset, RDMA from the whole population of conducted studies.

To model realistic conditions, the framework incorporates a simulated publication selection process that combines two mechanisms: (1) selective publication that favours significant results and (2) p-hacking that enables achieving the significance rates observed in the published literature (Figure 1). The relative contribution of each mechanism to the final rates of significant results in the published literature is recorded across simulation scenarios (see Section 3.2). Some primary studies achieve statistical significance naturally (i.e., when the true effect size and the sample size allow for this). When the number of such naturally significant studies is insufficient to meet the target sig-

nificance rates in the published literature, p-hacking is applied. Here, the smallest needed number of non-significant studies, whose natural significance is $0.05 < p \leq 0.30$, is manipulated to reach significance. This approach ensures that p-hacking is applied conservatively, only when needed to reach empirically observed proportions of published significant results. We impose a target proportion of significant results in the published literature rather than modelling publication selection as a probabilistic function of the p-value. Our main reason for this is that the target proportion is directly observable: the proportion of published studies that are statistically significant can be estimated from the literature [22, 23, 47, 48], whereas the probability that a study with a given p-value is published is largely unknown. In addition, fixing the target rate explicitly determines how much p-hacking each scenario requires, whereas a probabilistic selection model would add variance across iterations while leaving the expected composition of the published literature (which is what drives the distortion of CMA) unchanged.

The publication selection process yields $k = N/2$ published studies, and CMA synthesises all of them. We have chosen to sample 50% of conducted studies, as this reflects the publication rates observed across disciplines [18, 30, 49]. RDMA randomly samples $k = N/2$ studies from the population of all conducted studies (before the publication selection process). The performance of CMA and RDMA is then compared based on the accuracy of the effect size estimate, confidence interval coverage, and heterogeneity estimation (see Performance Metrics below for details). Compared to the logic of IPD meta-analysis in clinical research, our process is more relevant to other research areas (i.e., ecology, social sciences, environmental sciences) where effect sizes have to be calculated per study, rather than for pooled data across studies. In these fields, raw data cannot be pooled together because, compared to data from highly standardised clinical trials used in IPD meta-analyses, the pool of studies is heterogeneous. For example, studies use different methods, or are conducted on different species. We synthesise the raw-data arm with the same two-stage random-effects model as the conventional arm. Because both arms use the identical two-stage model, the only difference between them is the population of studies from which data are drawn, not the statistical machinery applied to synthesise them. Any advantage of RDMA we detect is therefore attributable to the data rather than to a more powerful synthesis method: in clinical IPD meta-analysis, a one-stage model, which fits a single model to the pooled participant-level data rather than combining study-level summaries, can be statistically more efficient, and such a model would be expected to do at least as well, though we do not test this directly; because both arms use the same two-stage model, our choice does not inflate RDMA's advantage.

For the ecological case study, the process is replicated 10,000 times for each combination of effect size, heterogeneity level, and significance rate ($3 \times 3 \times 2 = 18$ combinations, of which 15 were valid; see Section 2.2.3). The broader parameter sweep across publication and significance rates is described in Section 3.7.

2.2 Empirical parameter estimation

To understand how different conditions affect the performance of CMA and RDMA, we use ecology as a case study. However, many of the parameters we use are realistic across scientific disciplines (see Discussion). To obtain realistic simulation parameters, we used two ecology literature datasets: (1) Costello and Fox [47, 48], which compiled 467 ecological meta-analyses with over 111,000 effect sizes, and (2) Kimmel et al. [22, 23], which contains 26,747 statistical estimates from 350 ecology papers published in Ecology, Ecology Letters, Journal of Ecology, Science and Nature between January 2018 and May 2020. Based on the Costello and Fox [47, 48] dataset, we calculated the typical effect size magnitude, between-study heterogeneity, and the number of studies used in meta-analysis. Based on both datasets [22, 23, 47, 48], we calculated the prevalence of statistically significant results in published studies. Sample sizes of primary studies were derived from Fox [50] and Kimmel et al. [22, 23] (see Section 2.2.2 for details).

2.2.1 Effect sizes and heterogeneity

Based on Costello and Fox [47, 48], ecological meta-analyses included absolute Hedges' g values distributed with quartiles of 0.22 (Q_1), 0.55 (median), and 1.1 (Q_3). These empirical values align closely with Cohen's conventions [51] for small ($g = 0.2$), medium ($g = 0.5$), and large ($g = 0.8$) effect sizes. We therefore adopted Cohen's standardised thresholds for our simulations, both because they fall close to the empirical ecological quartiles and because they allow researchers from other fields to orientate on the same axis, consistent with our aim of a discipline-agnostic framework. For ecology, small- and medium-effect scenarios are the most representative, as ecological effect sizes are typically small to medium (median $|g| = 0.55$) computed from [47, 48].

We computed between-study heterogeneity (τ^2) from the ecological meta-analyses compiled by Costello and Fox [47, 48], finding a distribution with $Q_1 = 0.26$, median $= 0.53$, and $Q_3 = 0.88$. Critically, τ^2 showed a modest positive correlation with the mean magnitude of the meta-analytic effect size ($r = 0.41$, $p = 0.005$). To capture this empirical pattern, rather than using fixed τ^2 values across all effect sizes, we scaled heterogeneity proportionally to the effect size magnitude:

- Low heterogeneity: $\tau^2 = 0.40\times$ effect size

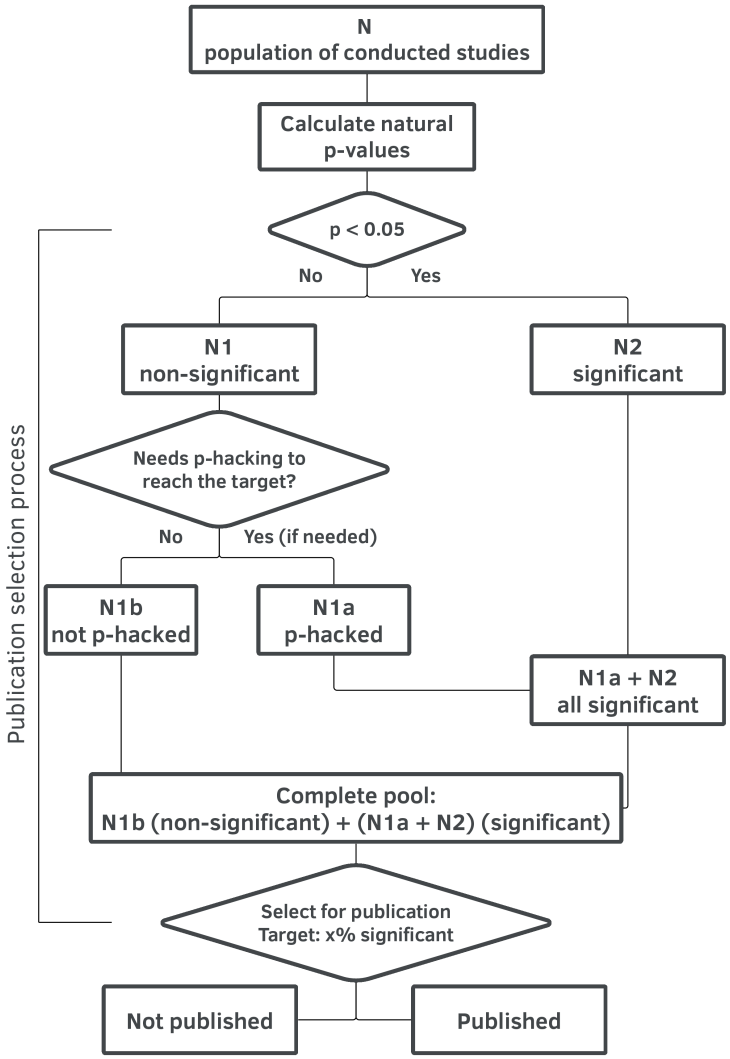


Figure 1: Flowchart illustrating the Classical meta-analysis (CMA) simulation process. The process begins with setting parameters: true effect size magnitude (θ), between-study heterogeneity (τ^2), sample sizes of primary studies (n_i), and population size of conducted studies (N). We then generate a population of N studies with known effect sizes and calculate natural p-values. Studies are then subjected to publication selection process with the aim to reach a target rate x% of $k = N/2$ published studies being significant. This process acts through selective publication and p-hacking.

- Medium heterogeneity: $\tau^2 = 1.00\times$ effect size
- High heterogeneity: $\tau^2 = 1.50\times$ effect size

This parametrisation produced τ^2 values that span from below the empirical Q_1 (0.08 for small effects with low heterogeneity) to above Q_3 (1.20 for large effects with high heterogeneity), covering the full empirical range while maintaining consistent relative heterogeneity across effect size magnitudes.

2.2.2 Meta-analysis and primary studies sample size

The median ecological meta-analysis comprises 64 effect sizes from 23 studies [47, 48]. Given that our simulation assumes one effect size per study, each meta-analysis in our simulation included 23 published studies (and thus effect sizes). This simplification means we do not capture within-study selective reporting or the statistical dependence among multiple effect sizes from the same study. This approach is conservative, since within-study selective reporting is a further source of distortion that affects published literature, but not raw-data synthesis with access to all measured outcomes. We further discuss the implications of this choice in the Discussion.

We determined the realistic sample size of the primary studies included in the meta-analysis using two existing large datasets, both of which have some limitations. The Fox [50] dataset provides sample sizes of studies included in ecological meta-analyses, but only for those that report correlation coefficients (median $n = 30$), and not for standardised mean differences (Hedges' g), which we use in our simulation. The Kimmel et al. [22, 23] dataset reports sample size for various effect sizes, including group comparisons suitable for Hedges' g (median $n = 79$). However, their studies come from the general primary literature rather than from studies included in meta-analyses. To reconcile these sources, we used a sampling approach that reflects the variation in the sample sizes of ecological studies: 25% of studies with $n = 20-30$ (small samples), 50% with $n = 30-70$ (moderate samples), and 25% with $n = 70 - 100$ (larger samples). This stratified sampling approach was applied to all $N = 46$ conducted primary studies in each simulation iteration, with sample sizes randomly drawn from the specified ranges. This yielded a median of around $n = 49$, a value that is intermediate between the median sample sizes of the primary studies included in the ecological meta-analysis ($n = 30$, [50]) and the median of the primary literature ($n = 79$, [22, 23]). The total sample size was split equally between the treatment and control groups.

Our approach captures low statistical power typical of many research fields. For example, Jennions and Møller [4] found only 13–16% power to detect small effects and 40–47% for medium effects in

behavioural ecology, reflecting the prevalent small sample sizes. Similarly, Cohen [6] reported mean power of just 48% for medium effects in psychology, Button et al. [7] estimated a median power of 21% across neuroscience, and Ioannidis et al. [8] found a median power of 18% or less across 159 areas of economics research.

2.2.3 Publication selection process rates

We characterise each scenario by its target significance rate, the percentage of published studies that contain a significant result, reached through both selective publication and p-hacking. Throughout, "publication bias" refers to the selective-publication process, the "publication rate" to the fraction of conducted studies that are published, and the "significance rate" to the percentage of significant results among published studies. Because each study contributes one effect size in our simulation, the percentage of significant studies equals the percentage of significant effects. Based on Kimmel et al. [22, 23] data, we calculated that 98.9% of published ecological studies contain at least one p-value below 0.05. From Costello and Fox [47, 48], we calculated that 64% of studies included in meta-analyses contain at least one significant result (amongst the results included in the meta-analysis). We use these two empirical proportions as the parameter values for our ecological case study: a 95% significance rate (near primary literature conditions, representing severe distortion) and a 65% significance rate (matching typical meta-analytic data, representing moderate distortion). One distinction matters for interpreting these values: the empirical proportions we take from the literature are study-level significance rates, that is, the proportion of studies reporting at least one significant result, whereas the effect-level significance rate that a meta-analyst actually samples is likely to be lower, because ecological studies typically report several effect sizes. For this reason we do not treat the two ecological values as fixed points but also explore the significance rate continuously between 30% and 95% (Figure 5), so that the framework covers scenarios from other fields, with the two ecological values retained as a case study within this range. For large effect sizes ($g = 0.8$), the natural significance rate exceeded 65% due to high statistical power (studies naturally achieved significance without bias), making the 65% scenario inapplicable. This resulted in 15 valid simulation scenarios for the ecological case study.

2.3 Data generation process

Step 1: Generate population of all conducted studies. We generated 46 studies representing all conducted research (i.e., before the publication selection process). We have chosen this number as it

is twice the number of studies used in meta-analysis in ecology (k=23, i.e., published studies in our case), and it reflects empirical data on how much of the conducted research is published. This is approximately 55% in ecology [30], 50% in clinical trials research [49], and 25–29% of doctoral dissertations in psychology [52]. Selective non-publication of null results is pervasive across the social sciences too (only 10 out of 48 null results were published, compared to 56 out of 91 strong results; [18]). This 2-to-1 ratio of conducted vs published studies determines how much p-hacking a given target significance rate implies.

For each study, we simultaneously: (1) drew one true effect size from $\theta_i = \mu_\theta + \delta_i$ where $\delta_i \sim N(0, \tau^2)$, using $\mu_\theta \in \{0.2, 0.5, 0.8\}$ and scaling τ^2 proportionally as described above; and (2) assigned a total sample size using stratified sampling (25% small: 20–30, 50% medium: 30–70, 25% large: 70–100), which were then split equally between the treatment and control groups.

Step 2: Simulate data and calculate effect sizes. For each study, we simulated raw data for treatment and control groups to generate the observed effect size (Hedges' g) and its sampling variance, following standard simulation approaches for meta-analysis [53, 54]. Each study's natural statistical significance was assessed with a two-sample Student t test on the simulated raw observations, as a primary study would report it, rather than derived from the standardised effect size; this distinction and its negligible numerical consequences are detailed in Supplementary Material S5.

Step 3: Create published literature through selective publication process. We constructed a pool of 23 published studies (the meta-analysis sample size) in which a target percentage of studies contained significant results. For the ecological case study this target was 95% or 65%; in the parameter sweep (Section 3.7) it ranged continuously from 30% to 95%. Some studies from the population of 46 conducted studies achieved statistical significance ($p < 0.05$) naturally in Step 2, based on their true effects and sample sizes. When the number of such naturally significant studies was insufficient to meet the target significance rate, we applied p-hacking to the minimum number of non-significant studies necessary to reach the target. We used only those studies with the original significance $0.05 < p \leq 0.30$, and ordered them by increasing p-value. This window encodes the assumption that only results already close to the conventional threshold can plausibly be moved across it through analytical flexibility; moving a result with $p = 0.6$ to significance would require manipulation closer to fabrication than to p-hacking. We report a sensitivity analysis widening this window to 0.40 and 0.50 in Supplementary Material S4. The studies selected for p-hacking are those closest to significance, drawn from the eligible pool with $0.05 < p \leq 0.30$ and ordered by increasing p-value. Of these selected studies, the half closest to significance

(those with the lowest p-values) underwent mild manipulation: the p-value was replaced with a random value drawn from a uniform distribution $U(0.01, 0.049)$ while retaining the original effect size. The remaining half of the selected studies (those with the higher p-values within this pool) underwent aggressive manipulation: a new p-value was drawn from $U(0.01, 0.049)$, a corresponding t-statistic was back-calculated from the new p-value and the study's degrees of freedom, and the effect size was recalculated as $g_i = t_i\sqrt{v_i}$, where v_i is the original sampling variance. Starting from those originally closest to statistical significance, only the minimum number of studies needed to reach the target significance rate were manipulated; all remaining studies retained their original p-values and effect sizes. This approach is intended to span the range of outcomes that questionable research practices produce, from a result nudged across the threshold to an inflated estimate [27]. Recent evidence suggests that more than half of statistically significant findings in environmental sciences may result from such practices [55].

Two features of this approach warrant justification. First, we model the net effect of p-hacking on the reported effect size and its p-value, rather than any specific analytical mechanism such as outlier removal, optional stopping, or switching between one- and two-tailed tests. We do so because no empirical study quantifies the relative prevalence of individual mechanisms or their quantitative effect on reported estimates, and because what a meta-analyst observes is only the reported effect size, its variance, and its p-value. Our two procedures are therefore intended to capture the possible outcomes of such practices rather than the practices themselves: moving a result across the significance threshold without changing the effect size (our mild manipulation), and inflating the estimate itself (our aggressive manipulation). We assign the two procedures by proximity to the threshold, on the assumption that a result already close to significance can be moved across it with minimal analytical intervention, leaving the estimate approximately intact, whereas a result further away requires practices that shift the estimate along with the p-value. We also conducted sensitivity analyses, reported in Supplementary Material S4. In particular, we vary the split between the two practices across its full range, from all manipulation only shifting the p-value across the threshold while leaving the effect size unchanged, to all manipulation inflating the effect size. The first extreme is the least favourable to RDMA, because leaving the effect size unchanged distorts CMA the least. Even there, RDMA remains more accurate than CMA, so the direction of our conclusion does not depend on the split assumed (see Supplementary Material S4, Figure 6). Second, the reported p-value and the reported effect size play different roles in our simulation. The p-value is used only when the published literature is assembled, to decide whether a study counts

towards the target proportion of significant results; the meta-analytic estimate itself is then computed only from the effect sizes and their variances, and never uses the p-value. As a result, when mild manipulation changes a study's p-value without changing its effect size, the two become mutually inconsistent, but this inconsistency affects only the assembly of the published set and never enters the meta-analytic estimate. This mirrors reality, where publication selection acts on reported significance while the synthesis is based on effect sizes. Holding the within-study variance fixed under p-hacking is a conservative simplification, because manipulation that also shrank the reported variance would give a study greater weight in the meta-analysis and would amplify, rather than reduce, the distortion of CMA. Our design therefore understates the advantage of RDMA rather than inflating it.

The resulting published literature of 23 studies that enters CMA combines naturally significant studies (randomly sampled from the available naturally significant studies), p-hacked studies, and a minority of non-significant studies (randomly sampled from the remaining non-significant studies) retained to reach the target proportion of significant results.

Step 4: Sample for meta-analysis. CMA uses all 23 studies from the published (biased) literature created in Step 3. RDMA randomly samples 23 studies from the original unbiased population of 46 conducted studies (Step 1, before p-hacking and publication selection). This choice makes two simplified assumptions: that CMA includes all available published studies (which is likely not true) and that raw data meta-analysis has access to the same number of studies. However, depending on the research question, RDMA might have access to more studies (i.e. where data are published, though the study was not) or to fewer studies (i.e. where datasets are not available). Our choice reflects a state between the ideal one (all raw data are available) and the worst one (no data available), while maintaining a balanced comparison by keeping the number of studies equal between the two methods, so that any detected differences are not attributable to a difference in sample size between CMA and RDMA.

Step 5: Meta-analytical estimates. Both CMA and RDMA use random-effects meta-analysis with REML estimation to obtain effect size estimates ($\hat{\theta}$), heterogeneity estimates ($\hat{\tau}^2$), and 95% confidence intervals. A complete mathematical formalisation of the simulation algorithm—including the exact conditioning structure, the construction of the published sample, and the expectation and bias of the CMA estimator—is provided in Supplementary Material S2.

2.4 Performance metrics

We evaluated the performance of CMA and RDMA using:

Mean absolute error (MAE): $\frac{1}{n}\sum_{i=1}^{n}|\hat{\theta}_i - \mu_\theta|$ is the average magnitude of the estimation error across all iterations, where $n = 10{,}000$ is the number of simulation iterations.

Relative mean absolute error (RelMAE): $\frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{\theta}_i-\mu_\theta|}{\mu_\theta}$ is the estimation error expressed as the proportion of the true effect size, enabling comparison across effect size magnitudes. Throughout this manuscript, when reporting percentage improvements or reductions between CMA and RDMA, we use this metric to enable fair comparison across different effect size magnitudes.

Confidence interval coverage: is the proportion of 95% CIs that contain the true effect size. Well-calibrated methods should achieve 95% or higher coverage.

Heterogeneity estimation accuracy: the mean absolute deviation $E[|\hat{\tau}-\tau|]$ and relative deviation $E[(\hat{\tau}-\tau)/\tau]$, measuring how well the between-study standard deviation ($\tau = \sqrt{\tau^2}$) is estimated.

2.5 Software

All analyses were conducted in R version 4.4.2 [56], using metafor (v4.6-0) [57], dplyr (v1.1.4) [58], tidyr (v1.3.1) [59], ggplot2 (v4.0.1) [60], and patchwork (v1.2.0) [61].

3 Results

3.1 Overview

Table 1 presents performance metrics across simulation scenarios. Overall, at a 95% significance rate, RDMA substantially outperformed CMA across the performance metrics. RDMA's advantages were most pronounced under low to medium heterogeneity. At a 65% significance rate, RDMA maintained advantages primarily for small to medium effects with low heterogeneity. Conceptually, these gains reflect the benefit of bypassing the publication filter, an advantage also relevant to IPD meta-analysis in clinical research.

3.2 P-hacking patterns

The simulation revealed that the contribution of p-hacking to the significance rate varied considerably with the effect size magnitude (mean p-hacking percentages are reported in Table 1). For small effects (ES=0.2) at a 95% significance rate, p-hacking rates also varied substantially across heterogeneity levels: 50.6% under low heterogeneity, 32.4% under medium heterogeneity, and 21.4% under high heterogeneity. For medium effects (ES=0.5), p-hacking was considerably lower

Table 1: Performance Metrics

Heterogeneity Level	Method	65% significance rate ME est.	MAE	RelMAE	CI Cov.	Mean p-hack %	95% significance rate ME est.	MAE	RelMAE	CI Cov.	Mean p-hack %
True ES = 0.2											
Low	CMA	0.37	0.17	0.85	0.58	23.02	0.48	0.28	1.41	0.21	50.59
	RDMA	0.20	0.07	0.34	0.94	—	0.21	0.07	0.33	0.95	—
Medium	CMA	0.33	0.15	0.75	0.82	6.94	0.44	0.24	1.22	0.57	32.43
	RDMA	0.20	0.09	0.44	0.93	—	0.20	0.09	0.44	0.93	—
High	CMA	0.30	0.15	0.75	0.88	2.12	0.41	0.23	1.13	0.71	21.41
	RDMA	0.20	0.10	0.51	0.94	—	0.20	0.10	0.52	0.93	—
True ES = 0.5											
Low	CMA	0.63	0.14	0.27	0.81	0.22	0.80	0.30	0.61	0.21	8.37
	RDMA	0.49	0.09	0.18	0.94	—	0.49	0.09	0.18	0.94	—
Medium	CMA	0.58	0.14	0.28	0.92	0.01	0.80	0.31	0.61	0.57	2.26
	RDMA	0.49	0.13	0.26	0.93	—	0.49	0.13	0.25	0.93	—
High	CMA	0.55	0.15	0.30	0.95	0.00	0.77	0.29	0.58	0.74	0.69
	RDMA	0.49	0.15	0.30	0.94	—	0.49	0.15	0.30	0.93	—
True ES = 0.8											
Low	CMA			—			1.07	0.27	0.33	0.34	0.08
	RDMA			—			0.78	0.11	0.13	0.93	—
Medium	CMA			—			1.10	0.31	0.39	0.64	0.04
	RDMA			—			0.78	0.16	0.20	0.94	—
High	CMA			—			1.08	0.31	0.39	0.77	0.01
	RDMA			—			0.79	0.19	0.23	0.93	—

Note: True ES = true effect size (Hedges' g). ME est. = Mean estimated effect size across 10,000 iterations. MAE = Mean Absolute Error. RelMAE = Relative Mean Absolute Error (MAE/true ES). CI Cov. = Confidence Interval Coverage (proportion of CIs containing the true ES; target: 0.95). Mean p-hack % = Mean percentage of published studies requiring p-hacking to reach the target significance rate across iterations. Heterogeneity levels are defined relative to effect size (τ^2 = 0.40×, 1.00×, 1.50× ES), so absolute τ^2 values differ across effect size scenarios (see Section 2.2.1)

(8.4%, 2.3%, and 0.7% for low, medium, and high heterogeneity, respectively), and for large effects (ES=0.8), essentially no p-hacking was needed (0.1%, 0.0%, and 0.0%), suggesting that selective non-publication dominated over questionable research practices in these scenarios. At a 65% significance rate, p-hacking was substantially reduced: small effects required 23.0%, 6.9%, and 2.1% p-hacking (low, medium, and high heterogeneity), while medium effects required minimal manipulation (0.2%, 0.0%, and 0.0%).

3.3 Effect size estimation across the scenarios

As expected from the simulation design, RDMA's estimates of the mean effect size are consistently centred on the true effect size (dashed lines), with distribution widths reflecting the inherent sampling variability from $k = 23$ studies and heterogeneity (Figure 2); the informative quantity is therefore not this property in itself but the magnitude of the deviation exhibited by CMA, and how it scales with the degree of publication selection. In contrast, CMA's estimates are systematically shifted upward at a 95% significance rate (Figure 2, Panel B), with the degree of the shift varying by effect size magnitude and heterogeneity level. RelMAE values (Figure 2, Panels C and D) demonstrate how these patterns translate to proportional estimation error.

The magnitude of CMA's overestimation directionally varied with the p-hacking rates (Section 3.2, Table 1). For small effects with low heterogeneity, and at a 95% significance rate, where p-hacking was the highest (50.6%), CMA deviated the most from the true effect, and RDMA achieved a 76% reduction in relative mean absolute error (RelMAE). This mechanistic link explained a counter-intuitive pattern visible in Table 1: for small effects, CMA's MAE decreases as heterogeneity increases ($0.28 \rightarrow 0.24 \rightarrow 0.23$). This occurred because increased heterogeneity reduced the amount of p-hacking implied by the target rate ($50.6\% \rightarrow 32.4\% \rightarrow 21.4\%$), thus reducing CMA's systematic bias, more than it increased the random error. For large effects with minimal p-hacking (0.1%), CMA's MAE followed the expected pattern, increasing with heterogeneity ($0.27 \rightarrow 0.31$). In contrast, RDMA's MAE increased monotonically with heterogeneity, confirming that errors reflect natural sampling variance rather than bias. For medium and large effects with minimal p-hacking (0.7–8.4% and 0.0–0.1%), CMA's overestimation remained substantial, demonstrating that selective publication alone distorts estimates.

At a 65% significance rate (Figure 2, Panel A), CMA's estimates are closer to true values than at 95%, yet systematic overestimation persisted, particularly for small effects with low heterogeneity, where p-hacking remained at 23.0%. For medium effects with high heterogeneity, where p-hacking was minimal ($\leq$0.2%), CMA and RDMA distribu-

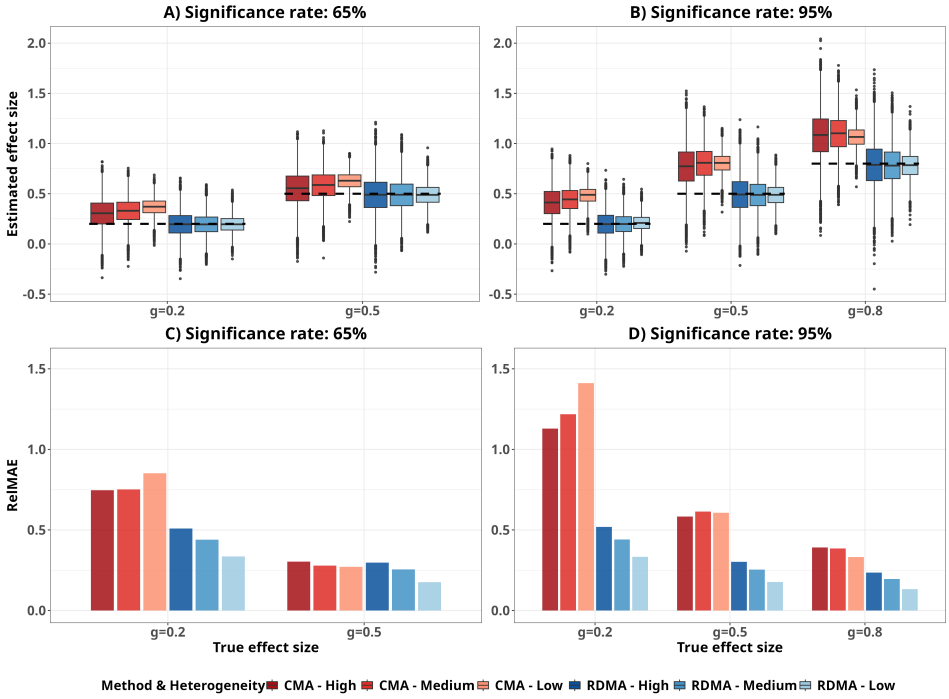


Figure 2: Effect size estimates and RelMAE comparing CMA and RDMA across scenarios. CMA is shown in red, RDMA in blue. The intensity of each colour represents the level of heterogeneity among the true effect sizes (darker colours represent higher heterogeneity). Panels A and B: Boxplots of estimated effect sizes at 65% and 95% significance rates, summarising 10,000 iterations with dashed horizontal lines indicating true effect sizes. Panels C and D: RelMAE at the same two significance rates. *Note: RDMA distributions are near-identical in panels A and B (and C and D) as this method samples from the pre-selection population; residual differences reflect Monte Carlo variation and the acceptance condition described in Supplementary Material S2. RDMA is displayed in both scenarios to facilitate direct visual comparison with CMA*

tions substantially overlap, indicating that CMA performs reasonably well in these specific conditions.

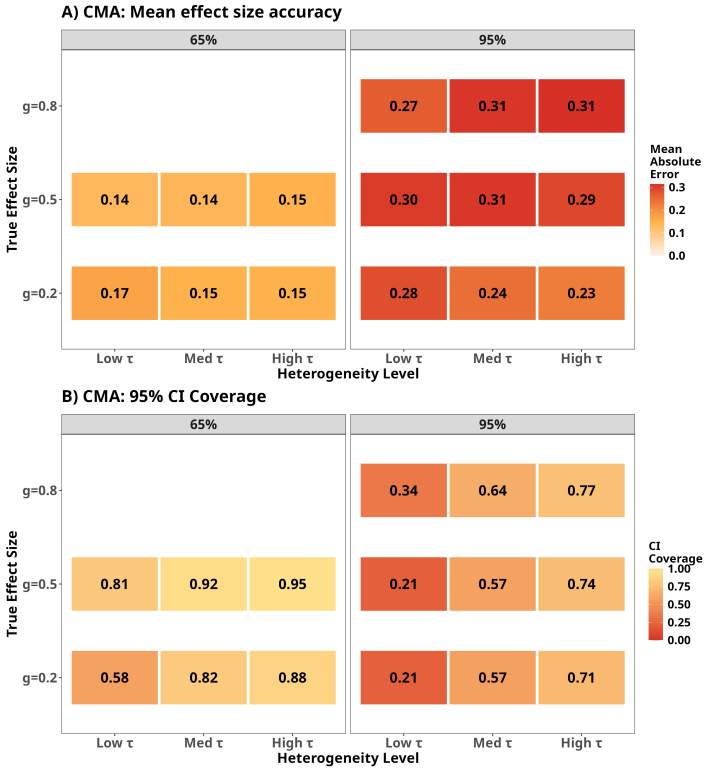


Figure 3: Performance heatmaps for CMA across simulation conditions. Panel A: Mean absolute error (MAE) in effect size estimates (lower values indicate greater accuracy). Panel B: 95% confidence interval coverage (target = 0.95). Rows represent true effect sizes ($g = 0.2, 0.5, 0.8$), columns represent heterogeneity levels (low, medium, high τ), and facets show the two significance rates (65% and 95% significant results amongst published results)

3.4 Confidence interval coverage

RDMA maintained near-nominal CI coverage (93–95%) across all simulation scenarios, closely approaching the 95% target expected of a well-calibrated method. CMA's CI coverage, in contrast, was severely degraded at a 95% significance rate, with the extent of coverage de-

pending critically on the effect size magnitude and heterogeneity (Figure 3, Panel B).

At a 95% significance rate, CMA's coverage was worst under low heterogeneity: only 21% coverage for small and medium effects (i.e. 79% of CIs failed to capture the true effect), and 34% for large effects. CMA performed slightly better under medium heterogeneity (57% coverage for small to medium effects, 64% for large effects), while high heterogeneity brought coverage closer to the acceptable levels (71–77%), though still substantially below RDMA's consistent 93–95%. This pattern is largely mechanical: higher heterogeneity inflates the estimated between-study variance, widening the confidence intervals and increasing the probability of capturing the true effect despite systematic bias in point estimates.

At a 65% significance rate, CMA approached acceptable performance for medium effects with medium-to-high heterogeneity (92–95% coverage), but still showed substantial undercoverage for small effects with low heterogeneity (58%).

3.5 Relative mean absolute error

RDMA exhibited minimal systematic bias (the mean estimates within 5% of true values across all scenarios). RDMA's RelMAE ranged from 13–52% depending on effect size and heterogeneity, primarily reflecting natural sampling variation from $k = 23$ studies rather than systematic over- or underestimation. In some scenarios, RDMA showed slight underestimation (e.g., mean estimate of 0.49 for true $ES = 0.5$), but RelMAE remained driven by the heterogeneity-induced variance. As heterogeneity increased within each magnitude of the effect size, RDMA's RelMAE increased proportionally (e.g., $ES = 0.2$: 34% $\rightarrow$ 44% $\rightarrow$ 51% across low, medium, and high heterogeneity), confirming that errors likely stem from sampling variance in heterogeneous populations rather than from bias.

In contrast, CMA showed both systematic overestimation and high RelMAE. For small effects with low heterogeneity at a 95% significance rate, CMA's mean estimate represented 141% overestimation with 141% RelMAE; nearly identical values indicating systematic bias rather than random variation. Standardising absolute error by effect size revealed different relative impacts: a similar MAE represents 141% error for $ES = 0.2$ but only 33% error for $ES = 0.8$, demonstrating why RelMAE is essential for comparing accuracy across effect sizes.

At a 65% significance rate, CMA's RelMAE improved substantially (27–85% across scenarios), approaching RDMA's performance for medium effects with high heterogeneity where p-hacking was minimal (Table 1, Figure 2).

3.6 Heterogeneity estimation

Both methods estimated heterogeneity reasonably well, with RDMA performing slightly better than CMA (Figure 4). RDMA's MAE for heterogeneity estimation (τ) was lower than CMA's in 13 of the 15 scenarios, with the largest advantage at a 95% significance rate: RDMA MAE 0.089 vs. CMA MAE 0.185, a difference of 0.096 units (ES = 0.2, high heterogeneity), and near-zero differences in the least biased scenarios. Complete heterogeneity estimation results for all the iterations are available in the accompanying data files.

At a 65% significance rate, the performance gap nearly disappeared for medium effects with medium-to-high heterogeneity, although RDMA maintained its advantage for small effects and low heterogeneity scenarios (Table 1, Figure 4).

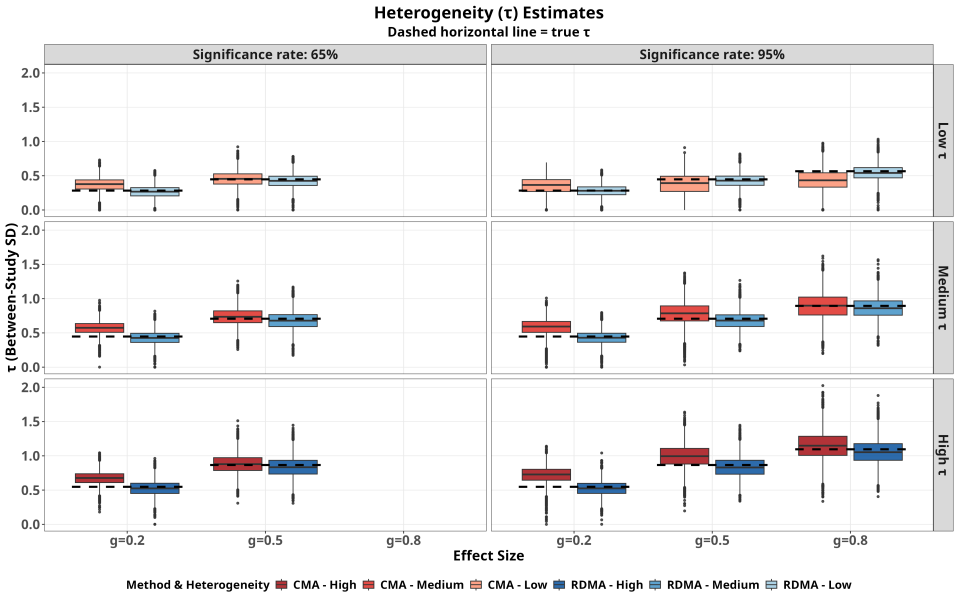


Figure 4: Heterogeneity estimates (τ, between-study standard deviation) from CMA (red) and RDMA (blue). Box-plots show distributions from 10,000 iterations. Dashed lines indicate true τ values under three scenarios of the true effect size values.

3.7 Robustness across publication rates and significance rates

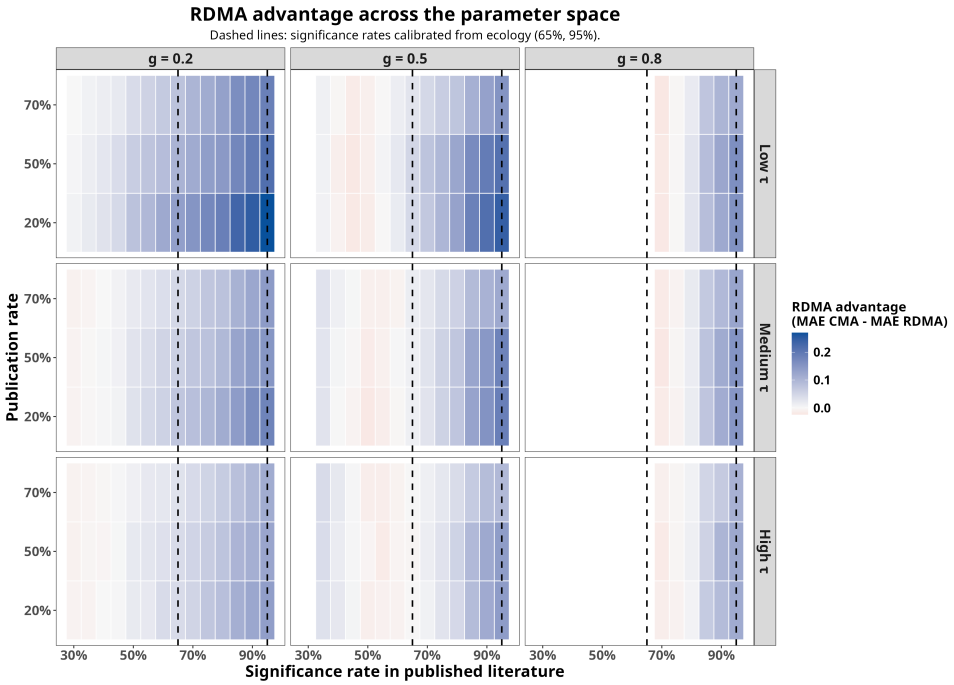


Figure 5: RDMA advantage (mean absolute error of CMA minus that of RDMA) across the publication rate and the significance rate in the published literature, for each effect size (columns) and heterogeneity level (rows). Blue indicates that RDMA is more accurate, red that CMA is. Dashed vertical lines mark the significance rates obtained from ecology (65% and 95%).

The results above use a 50% publication rate and the two significance rates obtained from ecology (65% and 95%). However, in other fields (or specific subareas of ecology) these rates might be different. Thus, we created additional scenarios with the publication rate set to 20%, 50% or 70%, and the significance rate in the published literature ranging from 30% to 95% in steps of 5%. Throughout, the number of published studies was kept at $k = 23$, so the size of the conducted population changed with the publication rate ($N \approx 115$, 46 and 33 at 20%, 50% and 70%), rather than being fixed at $N = 46$ as in the main analysis. As above, each combination was crossed with the three effect size magnitudes and three heterogeneity levels. We re-

tained only the combinations that describe an achievable published literature, that is, those whose target significance rate is no more than five percentage points below the natural significance rate for that effect size (a substantially lower target would describe a published literature less significant than the studies actually conducted, which is not a publication-bias scenario, and would eventually exhaust the supply of non-significant studies available to fill the remaining publication slots). This yielded 297 combinations, each run for 2,000 iterations. RDMA's advantage over CMA, measured as the difference in the mean absolute error (MAE of CMA minus MAE of RDMA), increased overall with the significance rate, from near zero at 30% to 0.043 at 65% and 0.156 at 95% on average, and grew as the publication rate fell, from 0.047 at a 70% publication rate to 0.066 at 20%.

Below a significance rate of roughly 50% the two methods were close on average, with the mean difference in mean absolute error staying below 0.015, because little selection or p-hacking is needed to assemble such a published literature. Two qualifications apply. In the small, homogeneous corner RDMA already holds an advantage at a 50% significance rate (0.081 at a 20% publication rate). Conversely, 64 of the 297 cells marginally favour CMA, by at most 0.024; these cluster where the target significance rate sits at or just below the natural rate, so the published literature is barely selected and the residual difference between the two arms is of the same order as the small downward bias (around 2%) that both show under our data-generating model. As the required significance rate rose, RDMA's advantage grew, and it was the largest for small, homogeneous effects when selection was most severe, that is, when a high significance rate had to be reached from a small published fraction. It peaked at a mean absolute error difference of 0.270 (CMA minus RDMA) for $g = 0.2$ under low heterogeneity, at a 95% significance rate and a 20% publication rate (Figure 5); the low publication rate leaves a large unpublished pool, so the published literature is a heavily selected slice of the conducted studies and CMA is correspondingly distorted. A separate corner of the parameter space is where the published literature is hardest to assemble: a high publication rate combined with a high significance rate. Here the unpublished pool is small, so almost every eligible study must be p-hacked to reach the target, and the required fraction of p-hacked studies rises steeply. Even so, every parameter combination could be assembled: only the single most extreme one ($g = 0.2$, low heterogeneity, 95% significance rate, 70% publication rate) ran out of eligible studies often enough to fall short of the full 2,000 iterations. The two ecological values (65% and 95%) sit within this landscape and are marked on the figure.

4 Discussion and Conclusions

We provide the first simulation-based quantification of the benefits that raw (open) data can bring to meta-analysis by mitigating the effects of publication bias and p-hacking, for the case in which raw data are available for the full population of conducted studies. Our results reveal that raw data meta-analysis yields widespread advantages over classical meta-analysis (when the sample number of included primary studies and their sample size is the same between the two). These advantages depend on the extent of publication bias, with the magnitude of the true effect size and the heterogeneity level both playing important moderating roles. Further, our simulations also revealed under what conditions the observed significance rate of the published literature can only be reached through p-hacking. We discuss our findings and their implications below.

4.1 Conditions requiring p-hacking to reach the observed significance rate

Our simulation revealed that, at the publication rate of roughly 50%, which is relevant across research fields, the contribution of p-hacking to the significance rate of the published literature largely depends on three factors: the magnitude of the true effect size, the heterogeneity between the studies, and the target significance rate; the publication rate itself is a fourth factor, explored in the parameter sweep (Section 3.7). The p-hacking implied by the observed significance rate is substantial when the true effects are small ($ES = 0.2$) and 95% of published results are significant, peaking at 50.6% under low heterogeneity. When heterogeneity of small effects is low, true effect sizes cluster tightly around a small mean value. Given that sample sizes typical for ecology (median $n \approx 49$) rarely achieve sufficient power to reach significance for such effects, the observed significance rate could only be reached through p-hacking. As heterogeneity of small effects increases, the distribution of true effect sizes widens, so some studies have larger true effects that achieve natural significance (statistical significance under the data-generating model), reducing the amount of p-hacking implied by the target significance rate. This explains why, paradoxically, CMA's mean absolute error decreases with increasing heterogeneity for $ES = 0.2$: higher heterogeneity reduces the proportion of studies requiring p-hacking, thereby reducing the severity of bias.

In contrast, at a 95% significance rate, minimal p-hacking was needed for medium effects (mean 3.8%) and essentially none for large effects (0.04%), where natural significance rates approach or exceed target levels due to adequate statistical power. Here, selective publica-

tion dominated over p-hacking. At a 65% significance rate, the implied p-hacking was substantially reduced but not eliminated: small effects still required considerable manipulation (23.0% under low heterogeneity), while medium effects required very little ($\leq$0.2%). These patterns reflect our modelling choice of an approximate 50% publication rate, which is likely realistic, and consistent with empirical evidence across fields [18, 30, 49, 52]. This ratio makes our estimates conservative: at lower publication rates the published literature is a more heavily selected subset of a larger population of conducted studies, so RDMA's advantage grows (Section 3.7), even though less p-hacking is needed to reach a given significance rate. Further, effects in ecology are typically small to medium (median $|g| = 0.55$; [47, 48]), and low statistical power is prevalent across many research fields [4, 7, 8]. Thus, our results demonstrate that the current incentives and publishing models, where significant results are favoured, are likely the key to pushing scientists to p-hack.

4.2 When to use raw data meta-analysis

Overall, the benefits of RDMA for accurately estimating the mean effect increase as true effect size and heterogeneity decrease, yet remain substantial in most scenarios. The benefits are also larger when 95% of published effects are statistically significant than when 65% are. These findings suggest that RDMA might provide the greatest value for research questions with relatively homogeneous effects, such as controlled experiments with standardised protocols, well-defined interventions, or phenomena governed by consistent mechanisms. It provides somewhat less value when studying highly context-dependent phenomena with inherently large heterogeneity, although it still yields meaningful improvements over CMA in these contexts.

RDMA's advantages are most pronounced for small effects with low heterogeneity (a 76% reduction in relative mean absolute error, RelMAE) because RDMA circumvents both selective publication and severe p-hacking in these scenarios. For larger effects, RDMA primarily circumvents selective publication alone (as p-hacking rates in this scenario are small), yielding a smaller but still substantial advantage (a 40–60% reduction in RelMAE). Higher heterogeneity reduces RDMA's relative advantage across all effect sizes, measured as RelMAE, because true variation among studies becomes large relative to the distortion introduced by bias. Critically, the type of error differs between the two methods. RDMA's errors stem from natural sampling variation: individual meta-analyses vary around the true effect, as expected when randomly sampling $k = 23$ studies from a larger population. This produces unbiased estimates with moderate precision. CMA's errors, in contrast, reflect systematic bias, with estimates consistently shifted

away from true values in a predictable direction. This distinction is fundamental: random error decreases with more studies, while systematic bias persists regardless of sample size. RDMA therefore provides not just more accurate estimates, but estimates that improve predictably as data accumulate.

Our simulation suggests that RDMA's primary advantage lies in accessing the unbiased population of conducted studies, bypassing the publication selection filter that distorts CMA estimates. In our framework, both RDMA and CMA apply the same analytical method (random-effects meta-analysis of study-level effect sizes). The difference in performance therefore comes entirely from which studies each method draws on: RDMA uses the effect sizes of all conducted studies, computed from raw data and unaffected by selective publication or p-hacking, whereas CMA uses the effect sizes reported in the published literature, which may be distorted by both. Our results therefore quantify the access channel through which raw data benefits meta-analysis. This advantage may be shared by other approaches that bypass the publication filter and use original datasets, such as IPD meta-analysis in clinical research, though the mechanisms differ. For simple treatment–control comparisons with continuous outcomes—the design that our simulation captures—published evidence suggests that one-stage participant-level modelling and two-stage study-level approaches yield practically equivalent summary treatment effects [62–64], though our simulation does not test this directly. The additional analytical advantages of IPD meta-analysis—such as consistent covariate adjustment, exploration of participant-level treatment-effect modifiers, and appropriate handling of time-to-event outcomes—become important for more complex research questions that our framework does not address.

4.3 Limitations and future directions

We make several simplifying assumptions in our simulations. First, we conservatively modelled p-hacking only when needed to reach target significance rates, making our RDMA benefit estimates conservative; in reality, questionable practices are likely more diverse. Second, we assumed that RDMA draws its sample from the full population of conducted studies with complete data, whether or not they were published, taking the same number of studies as CMA. This does not mean RDMA uses every conducted study; like CMA, it synthesises a sample of 23. What it assumes is that the pool RDMA samples from is unbiased, containing published and unpublished studies alike and using their original, unmanipulated effect sizes, rather than the publication-filtered pool available to CMA. In practice, data from all conducted studies (published or unpublished) will not be available or of sufficient

quality for use [43], reducing RDMA's advantages. Thus, our results represent an upper bound of the benefits. Access to raw data also does not resolve every problem in synthesis: studies whose data are never archived remain in the file drawer, poor experimental design is not repaired by reanalysis, and inadequate annotation or missing metadata can render archived data unusable. Raw data improve inference; they do not guarantee it. Similar constraints are well recognised in IPD meta-analysis, where data availability and access decisions can shape the final evidence base. However, RDMA can draw on more studies than CMA when published literature is limited [41]. On the other hand, in an ideal scenario where open data are the norm, RDMA might have access to a larger pool of studies than CMA, likely increasing the benefits beyond those reported here.

Third, our simulation examined only a single effect size per study from simple treatment-control comparisons and therefore did not capture potential within-study heterogeneity in effect sizes for studies that report on more than one effect. Related to this, our representation of the publication selection process enforced a target prevalence of statistically significant results among published studies, with one effect size per study. In practice, selection mechanisms are more complex across disciplines: entire studies may remain unpublished (file drawer effect), researchers may selectively report significant results while omitting non-significant ones from the same study, and published studies typically contain multiple effect sizes rather than one. Some research areas, such as clinical research, may exhibit additional structured selection mechanisms, including time-lag bias and sponsor-dependent dissemination patterns. Fourth, we tested only random-effects models with REML estimation; other meta-analytic approaches might show different patterns. Finally, we did not examine meta-regression or subgroup analyses, which are common in practice.

Future work should thus examine scenarios with multiple effects per study to understand how within-study selection interacts with publication bias. Future extensions could also implement explicit selection probability models, in which the probability of publication is modelled as a function of statistical significance (e.g., high probability for $p < 0.05$ and lower probability otherwise) and study characteristics, including sponsor status. Such models are widely used in clinical selection-model literature and would allow direct comparison with aggregate-data correction approaches developed for medical meta-analysis. Incorporating sponsor-dependent or outcome-level selection would be particularly relevant for extending our framework to broader bias scenarios encountered in regulatory or guideline-setting contexts.

Finally, our choice of 50% publication rate was based on empirical evidence from ecology and other fields [18, 30, 49, 52], and represents a conservative scenario. Our 95% significance rate scenario may be

less realistic for ecology, where meta-analyses typically include studies that report multiple effect sizes, and only 64% of such studies contain at least one significant result among the effects included in the meta-analysis [47, 48]. However, this scenario may be more realistic in fields where primary studies commonly report a single effect size, which is then more likely to be significant. We also varied publication rates between 20% and 70% (Section 3.7, Figure 5, Supplementary Material S6) and the significance rate in the published literature between 30% and 95%. When the publication rate is lower (for example 20%, requiring $N \approx 115$ conducted studies for $k = 23$ published), the larger unpublished pool allows the target significance rate to be reached largely through publication selection, so the published literature is a more heavily selected subset of the conducted studies and RDMA's advantage grows. When publication rates are higher (for example 70%, requiring only $N \approx 33$ conducted studies), the advantage is smaller. The analysis also shows that the implied fraction of p-hacked studies rises steeply towards the extreme corner (small effects, high significance rate, high publication rate), where a valid published literature becomes progressively harder to construct, which further clarifies the boundary conditions for RDMA's benefits.

Our estimates cannot at the moment be evaluated against real data as raw data meta-analysis remains very rare. However, if more meta-analyses begin using raw data, following existing guidelines [36], it might be possible to evaluate the assumptions of our model and its estimates. Hybrid approaches that combine published effect sizes with available raw data merit investigation, as they may provide a pragmatic middle ground. Empirical studies quantifying the time and resource costs of data acquisition would help researchers make informed decisions about when to pursue RDMA. Such extensions would strengthen understanding of when RDMA justifies the additional time investment it requires.

4.4 Tool for meta-analysts

Our simulation framework is designed to be field-agnostic. Researchers can assess RDMA's potential benefits in their domain (or a particular research area) by replacing our ecological parameters with values relevant to their scenario. Required inputs include: (1) expected (typical) effect size magnitudes (Hedges' g or other standardised effect sizes), (2) between-study heterogeneity estimates (τ^2), (3) typical sample sizes of primary studies (e.g., treatment and control group sizes), (4) the typical number of published studies per meta-analysis ($k = 23$ in our simulations, representing the median in ecology), (5) the estimated publication rate in the field (approximately 50% in ecology, meaning $N = 46$ conducted studies are needed to yield $k = 23$ published

studies), and (6) the target significance rate (proportion of significant results in the published literature). Both k and the publication rate are necessary because together they determine N, the population of conducted studies from which RDMA samples, and consequently how strongly the published literature is selected and how much p-hacking is required to reach the target significance rate: a lower publication rate (requiring a larger N of conducted studies to yield k published studies) allows stronger selection with less p-hacking and amplifies RDMA's advantage, whereas a higher publication rate requires more p-hacking.

The code, which can be accessed at [45], automatically generates performance metrics (mean absolute error, confidence interval coverage, heterogeneity estimation accuracy) across user-specified scenarios. For researchers who want a quick preliminary estimate of expected CMA bias without running the full simulation, we also provide a closed-form approximation based on truncated normal theory (Supplementary Material S3) with an accompanying R script (`S3.R`) that returns an approximate bias estimate for a given combination of effect size, heterogeneity, sample sizes, and target significance rate. Although our simulation used Hedges' g as the effect size measure, the framework can be adapted to other common meta-analytic metrics such as the log response ratio (lnRR) or Fisher's z-transformed correlation coefficient. The overall architecture—including the publication bias mechanism, the RDMA versus CMA sampling design, and all performance metrics—is effect size agnostic. Adaptation requires replacing the data-generating function with one appropriate for the targeted effect size, adjusting the p-hacking mechanism to match the corresponding test statistic, and recalibrating input parameters to reflect empirical distributions for the chosen metric. All manipulation is confined to a single block of the literature-generation function, documented in the code repository, so a reader who wishes to model a specific mechanism (for example outlier removal, optional stopping, or selective outcome switching) rather than our net-effect formulation can substitute it there without altering the rest of the pipeline. Detailed guidance is provided in the code repository [45].

4.5 Conclusions

RDMA is overall superior to CMA; however, its advantages depend on the interaction of the significance rate in the published literature, the publication rate, the true effect size magnitude, and the heterogeneity level. The advantage we quantify arises from access to an unbiased, unselected set of conducted studies, that is, from bypassing publication selection and the incentive to p-hack. It does not arise from the synthesis method, since both methods apply the same two-stage random-effects model. Where raw data are available only for

a selectively published subset of studies, the gains reported here are expected to be lower. When 95% of published results are significant, RDMA substantially reduces relative mean absolute error compared to CMA: by 54–76% for small effects ($g = 0.2$), 48–71% for medium effects ($g = 0.5$), and 40–60% for large effects ($g = 0.8$). RDMA consistently maintained 93–95% confidence interval coverage across all scenarios, while CMA's coverage ranged from only 21–34% under low heterogeneity to 71–77% under high heterogeneity. At a 65% significance rate with high heterogeneity and medium effects, the performance gap narrows substantially, with CMA performing adequately.

Given that raw data meta-analysis provides substantially more accurate effect size estimates across the scenarios most plausible for a range of research fields, researchers should more often seek to incorporate raw data into meta-analyses. This can be done by either using raw data from published studies or by identifying raw data independently [36, 41]. However, raw data are often difficult to find and use, as their quality is often very low [43, 44]. The research community, including funders and publishers, should thus seek to increase data quality and thereby increase the reliability of estimated effects.

Author contributions: Conceptualisation: D.Ž., A.C.; Methodology: D.Ž., S.G., A.T., A.C.; Software: D.Ž.; Validation: S.G., A.T.; Formal analysis: D.Ž.; Visualisation: D.Ž.; Funding acquisition: A.C.; Writing—original draft: D.Ž., A.C.; Writing—review and editing: S.G., A.T., A.C., L.P.

Conflict of interest disclosure: The authors of this preprint declare that they have no financial conflict of interest with the content of this article. The authors declare they have no conflict of interest relating to the content of this article.

Funding: This work was supported by the Croatian Science Foundation under the project number HRZZ-IP-2022-10-2872.

Data, script and code availability: All simulation code, parameter estimation scripts, input data, and complete results are available in the OSF repository: https://doi.org/10.17605/OSF.IO/VEBYM [45].

Acknowledgements: We are grateful to Dr Alfredo Sánchez-Tójar for providing feedback on aspects of the methodology. AI tools (Claude, Anthropic) were used to assist with writing the code and editing the manuscript. All code and text were checked by the authors, who take full responsibility for the content.

Ethics: This study is based entirely on simulated data. No ethical approval was required.

References

[1] S. Nakagawa et al. "Quantitative evidence synthesis: A practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences". In: *Environmental Evidence* 12.1 (2023), p. 8. DOI: 10.1186/s13750-023-00301-6.

[2] M. Borenstein et al. *Introduction to Meta-Analysis*. 2nd. Wiley, 2021. ISBN: 978-1-119-55835-4.

[3] J.P.T. Higgins et al. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd. Wiley-Blackwell, 2019. ISBN: 978-1-119-53662-8.

[4] M.D. Jennions and A.P. Møller. "A survey of the statistical power of research in behavioral ecology and animal behavior". In: *Behavioral Ecology* 14.3 (2003), pp. 438–445. DOI: 10.1093/beheco/14.3.438.

[5] Y. Yang et al. "Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology". In: *Global Change Biology* 28.3 (2022), pp. 969–989. DOI: 10.1111/gcb.15972.

[6] J. Cohen. "The statistical power of abnormal-social psychological research: A review". In: *The Journal of Abnormal and Social Psychology* 65.3 (1962), pp. 145–153. DOI: 10.1037/h0045186.

[7] K.S. Button et al. "Power failure: why small sample size undermines the reliability of neuroscience". In: *Nature Reviews Neuroscience* 14.5 (2013), pp. 365–376. DOI: 10.1038/nrn3475.

[8] J.P.A. Ioannidis, T.D. Stanley, and H. Doucouliagos. "The power of bias in economics research". In: *The Economic Journal* 127.605 (2017), F236–F265. DOI: 10.1111/ecoj.12461.

[9] A. Brodeur, N. Cook, and A. Heyes. "Methods matter: p-hacking and publication bias in causal analysis in economics". In: *American Economic Review* 110.11 (2020), pp. 3634–3660. DOI: 10.1257/aer.20190687.

[10] S.B. Bruns and J.P.A. Ioannidis. "p-Curve and p-hacking in observational research". In: *PLOS ONE* 11.2 (2016), e0149144. DOI: 10.1371/journal.pone.0149144.

[11] M.L. Head et al. "The extent and consequences of p-hacking in science". In: *PLOS Biology* 13.3 (2015), e1002106. DOI: 10.1371/journal.pbio.1002106.

[12] D. Fanelli. "'Positive' results increase down the hierarchy of the sciences". In: *PLOS ONE* 5.4 (2010), e10068. DOI: 10.1371/journal.pone.0010068.

[13] D. Fanelli. "Negative results are disappearing from most disciplines and countries". In: *Scientometrics* 90.3 (2012), pp. 891–904. DOI: 10.1007/s11192-011-0494-7.

[14] T.D. Sterling. "Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa". In: *Journal of the American Statistical Association* 54.285 (1959), pp. 30–34. DOI: 10.1080/01621459.1959.10501497.

[15] Y. Yang et al. "Publication bias impacts on effect size, statistical power, and magnitude (Type M) and sign (Type S) errors in ecology and evolutionary biology". In: *BMC Biology* 21.1 (2023), p. 71. DOI: 10.1186/s12915-022-01485-y.

[16] S. Nakagawa et al. "Methods for testing publication bias in ecological and evolutionary meta-analyses". In: *Methods in Ecology and Evolution* 13.1 (2022), pp. 4–21. DOI: 10.1111/2041-210X.13724.

[17] E.H. Turner et al. "Selective publication of antidepressant trials and its influence on apparent efficacy". In: *The New England Journal of Medicine* 358.3 (2008), pp. 252–260. DOI: 10.1056/NEJMsa065779.

[18] A. Franco, N. Malhotra, and G. Simonovits. "Publication bias in the social sciences: unlocking the file drawer". In: *Science* 345.6203 (2014), pp. 1502–1505. DOI: 10.1126/science.1255484.

[19] J.P. Simmons, L.D. Nelson, and U. Simonsohn. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant". In: *Psychological Science* 22.11 (2011), pp. 1359–1366. DOI: 10.1177/0956797611417632.

[20] I.A. Cristea and J.P.A. Ioannidis. "P values in display items are ubiquitous and almost invariably significant: A survey of top science journals". In: *PLOS ONE* 13.5 (2018), e0197440. DOI: 10.1371/journal.pone.0197440.

[21] D. Chavalarias et al. "Evolution of reporting p values in the biomedical literature, 1990-2015". In: *JAMA* 315.11 (2016), pp. 1141–1148. DOI: 10.1001/jama.2016.1952.

[22] K. Kimmel, M.L. Avolio, and P.J. Ferraro. "Empirical evidence of widespread exaggeration bias and selective reporting in ecology". In: *Nature Ecology & Evolution* 7.9 (2023), pp. 1525–1536. DOI: 10.1038/s41559-023-02144-3.

[23] K. Kimmel and P.J. Ferraro. *Replicability in ecology*. Dataset, OSF. Accessed 15 December 2025. 2023. URL: https://osf.io/9yd2b.

[24] L.K. John, G. Loewenstein, and D. Prelec. "Measuring the prevalence of questionable research practices with incentives for truth telling". In: *Psychological Science* 23.5 (2012), pp. 524–532. DOI: 10.1177/0956797611430953.

[25] F. Agnoli et al. "Questionable research practices among Italian research psychologists". In: *PLOS ONE* 12.3 (2017), e0172792. DOI: 10.1371/journal.pone.0172792.

[26] K. Fiedler and N. Schwarz. "Questionable research practices revisited". In: *Social Psychological and Personality Science* 7.1 (2016), pp. 45–52. DOI: 10.1177/1948550615612150.

[27] H. Fraser et al. "Questionable research practices in ecology and evolution". In: *PLOS ONE* 13.7 (2018), e0200303. DOI: 10.1371/journal.pone.0200303.

[28] A.A. Baptista and F. Pereira. "Questionable research practices in engineering research". In: *International Journal of Information Systems and Project Management* 14.1 (2026), pp. 1–24. DOI: 10.12821/ijispm140101.

[29] G. Gopalakrishna et al. "Prevalence of questionable research practices, research misconduct and their potential explanatory factors: a survey among academic researchers in The Netherlands". In: *PLOS ONE* 17.2 (2022), e0263023. DOI: 10.1371/journal.pone.0263023.

[30] M. Purgar, T. Klanjscek, and A. Culina. "Quantifying research waste in ecology". In: *Nature Ecology & Evolution* 6.9 (2022), pp. 1390–1397. DOI: 10.1038/s41559-022-01820-0.

[31] M. Friese and J. Frankenbach. "p-Hacking and publication bias interact to distort meta-analytic effect size estimates". In: *Psychological Methods* 25.4 (2020), pp. 456–471. DOI: 10.1037/met0000246.

[32] A.M. Senior et al. "Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications". In: *Ecology* 97.12 (2016), pp. 3293–3299. DOI: 10.1002/ecy.1591.

[33] S. van Erp et al. "Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990–2013". In: *Journal of Open Psychology Data* 5 (2017), p. 4. DOI: 10.5334/jopd.33.

[34] J.P. Higgins et al. "Measuring inconsistency in meta-analyses". In: *BMJ* 327.7414 (2003), pp. 557–560. DOI: 10.1136/bmj.327.7414.557.

[35] Digital Science et al. *The State of Open Data 2023*. 2023. DOI: 10.6084/m9.figshare.24428194.v1. URL: https://digitalscience.figshare.com/articles/report/The_State_of_Open_Data_2023/24428194.

[36] A. Culina et al. "How to do meta-analysis of open datasets". In: *Nature Ecology & Evolution* 2.7 (2018), pp. 1053–1056. DOI: 10.1038/s41559-018-0579-2.

[37] E.T. Borer et al. "Finding generality in ecology: a model for globally distributed experiments". In: *Methods in Ecology and Evolution* 5.1 (2014), pp. 65–73. DOI: 10.1111/2041-210X.12125.

[38] B.L. Houtkoop et al. “Data sharing in psychology: A survey on barriers and preconditions”. In: *Advances in Methods and Practices in Psychological Science* 1.1 (2018), pp. 70–85. DOI: 10.1177/2515245917751886.

[39] A. Pepe et al. “How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers”. In: *PLOS ONE* 9.8 (2014), e104798. DOI: 10.1371/journal.pone.0104798.

[40] G. Colavizza et al. “The citation advantage of linking publications to research data”. In: *PLOS ONE* 15.4 (2020), e0230416. DOI: 10.1371/journal.pone.0230416.

[41] A. Culina et al. “Navigating the unfolding open data landscape in ecology and evolution”. In: *Nature Ecology & Evolution* 2.3 (2018), pp. 420–426. DOI: 10.1038/s41559-017-0458-2.

[42] L. Tedersoo et al. “Data sharing practices and data availability upon request differ across scientific disciplines”. In: *Scientific Data* 8 (2021), p. 192. DOI: 10.1038/s41597-021-00981-0.

[43] D.G. Roche et al. “Public data archiving in ecology and evolution: how well are we doing?” In: *PLOS Biology* 13.11 (2015), e1002295. DOI: 10.1371/journal.pbio.1002295.

[44] D.G. Roche et al. “Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution”. In: *Proceedings of the Royal Society B: Biological Sciences* 289.1975 (2022), p. 20212780. DOI: 10.1098/rspb.2021.2780.

[45] D. Žanko. *More reliable inference from open raw data: a Monte Carlo comparison of raw-data and aggregate-data meta-analysis under publication bias.* OSF. Accessed 27 July 2026. 2026. URL: https://doi.org/10.17605/OSF.IO/VEBYM.

[46] D. Žanko et al. *Performance of open data versus classical meta-analysis in ecology - a simulation study.* OSF Registries. Accessed 15 December 2025. 2025. URL: https://doi.org/10.17605/OSF.IO/KAH42.

[47] L. Costello and J.W. Fox. "Decline effects are rare in ecology". In: *Ecology* 103.6 (2022), e3680. DOI: 10.1002/ecy.3680.

[48] J.W. Fox and L. Costello. *Decline effects are rare in ecology: a meta-meta-analysis*. Accessed 15 December 2025. Dryad, 2022. URL: https://doi.org/10.5061/dryad.zkh1893b7.

[49] I. Chalmers and P. Glasziou. "Avoidable waste in the production and reporting of research evidence". In: *Lancet* 374.9683 (2009), pp. 86–89. DOI: 10.1016/S0140-6736(09)60329-9.

[50] J.W. Fox. "The sample size of the typical ecological correlation coefficient is small and slowly declining". In: *Oikos* 2025.12 (2025), e11430. DOI: 10.1002/oik.11430.

[51] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. 2nd. Routledge, 1988. ISBN: 978-0805802832.

[52] S.C. Evans et al. "'Are you gonna publish that?' Peer-reviewed publication outcomes of doctoral dissertations in psychology". In: *PLOS ONE* 13.2 (2018), e0192219. DOI: 10.1371/journal.pone.0192219.

[53] F. Gambarota and G. Altoè. "Understanding meta-analysis through data simulation with applications to power analysis". In: *Advances in Methods and Practices in Psychological Science* 7.1 (2024). DOI: 10.1177/25152459231209330.

[54] L.V. Hedges. "Distribution theory for Glass's estimator of effect size and related estimators". In: *Journal of Educational Statistics* 6.2 (1981), pp. 107–128. DOI: 10.3102/10769986006002107.

[55] T. Deressa et al. *More Than Half of Statistically Significant Research Findings in the Environmental Sciences are Actually Not*. Accessed 15 December 2025. 2023. URL: https://doi.org/10.32942/x24g6z.

[56] R Core Team. *R: A language and environment for statistical computing*. Accessed 15 December 2025. R Foundation for Statistical Computing. Vienna, Austria, 2024. URL: https://www.R-project.org/.

[57] W. Viechtbauer. "Conducting meta-analyses in R with the metafor package". In: *Journal of Statistical Software* 36.3 (2010), pp. 1–48. DOI: 10.18637/jss.v036.i03.

[58] H. Wickham et al. *dplyr: A grammar of data manipulation*. R package version 1.1.4. Accessed 15 December 2025. 2023. URL: https://CRAN.R-project.org/package=dplyr.

[59] H. Wickham, D. Vaughan, and M. Girlich. *tidyr: Tidy messy data*. R package version 1.3.1. Accessed 15 December 2025. 2024. URL: https://CRAN.R-project.org/package=tidyr.

[60] H. Wickham. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag, 2016. ISBN: 978-3-319-24277-4.

[61] T.L. Pedersen. *patchwork: The composer of plots*. R package version 1.2.0. Accessed 15 December 2025. 2024. URL: https://CRAN.R-project.org/package=patchwork.

[62] D.L. Burke, J. Ensor, and R.D. Riley. "Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ". In: *Statistics in Medicine* 36.5 (2017), pp. 855–875. DOI: 10.1002/sim.7141.

[63] R.D. Riley et al. "Two-stage or not two-stage? That is the question for IPD meta-analysis projects". In: *Research Synthesis Methods* 14.6 (2023), pp. 903–910. DOI: 10.1002/jrsm.1661.

[64] T.P. Morris et al. "Meta-analysis of Gaussian individual patient data: two-stage or not two-stage?" In: *Statistics in Medicine* 37.9 (2018), pp. 1419–1438. DOI: 10.1002/sim.7589.

Supplementary Material S1: Deviations from Preregistration

Our simulation approach evolved during empirical data analysis to better reflect realistic conditions in ecological meta-analysis. Key modifications align simulation parameters with empirical evidence while maintaining core objectives.

Number of studies per meta-analysis

We initially planned $k = 70$ studies based on the median number of effect sizes per meta-analysis. However, examining Fox and Costello [48] more carefully revealed that the median meta-analysis comprises 64 effect sizes from only 23 unique studies. Since our simulation generates one effect size per study, $k = 23$ better reflects typical primary studies in ecological meta-analyses.

Sample size allocation

Our preregistration specified 50% of studies with sample sizes $n = 20-30$, 25% with $n = 30-70$, and 25% with $n = 70-100$. We revised this to 25% with $n = 20-30$, 50% with $n = 30-70$, and 25% with $n = 70-100$ to better represent the range of sample sizes in ecological research while maintaining prevalence of small samples.

Heterogeneity parametrisation

We initially specified heterogeneity as $\tau = 10-30\%$ of effect size (τ being standard deviation). Empirical analysis revealed modest positive correlation between absolute effect size and τ^2 (variance) ($r = 0.41$, $p = 0.005$), leading us to express heterogeneity as proportions of effect size magnitude: $\tau^2 = 0.40\times$ effect size (low), $1.00\times$ effect size (medium), and $1.50\times$ effect size (high). This produces values spanning from below empirical Q_1 (0.26) to above Q_3 (0.88) observed by Fox and Costello [48].

Additional significance rate scenario

While our preregistration focused on a 95% significance rate, we added a 65% significance rate scenario based on Fox and Costello [48], which showed 64% of studies in ecological meta-analyses contain significant results.

These refinements improve our simulation's ecological realism without changing fundamental research questions or analytical approach.

Additional analyses added during revision

In response to peer review we added three analyses that were not part of the original preregistration: (i) a continuous sweep of the publication rate (20% to 70%) and the significance rate in the published literature (30% to 95%), reported in the main text and Supplementary Material S6; (ii) sensitivity analyses varying the p-hacking eligibility window and the split between threshold-shifting and estimate-inflating manipulation, reported in Supplementary Material S4; and (iii) a change in how each study's natural significance is computed, now from a two-sample t test on the simulated raw data rather than from the standardised effect size, with the two definitions compared in Supplementary Material S5. These are additions and a code correction rather than departures from the preregistered design, and none alters the preregistered research questions or analytical approach.

Supplementary Material S2: Mathematical formalisation of the simulation algorithm

Here we formalise the publication-bias and p-hacking mechanism exactly as implemented in the simulation code. We explicitly state the conditions under which the algorithm is successful, and calculate the expected total reported effect under those conditions.

Exact conditioning structure induced by the simulation algorithm

Let k denote the target number of published studies entering the classical meta-analysis, and let K denote the size of the conducted-study population generated in each attempt ($k = 23$; $K = 46$ in the main analysis, while in the parameter sweep K equals k divided by the publication rate, rounded to the nearest integer). The code repeats generation attempts until a valid published literature is obtained, so all expectations relevant to the simulation are conditional on the event that the generation algorithm succeeds.

For each conducted study $i = 1, \ldots, K$, a total sample size N_i is generated, then split as

$$n_{Ti} = \left\lceil \frac{N_i}{2} \right\rceil, \qquad n_{Ci} = \left\lfloor \frac{N_i}{2} \right\rfloor, \qquad \mathrm{df}_i = N_i - 2.$$

A true study-specific effect is then generated as

$$\Delta_i = \mu + U_i, \qquad U_i \sim \mathcal{N}(0, \tau^2).$$

Conditional on (N_i, Δ_i), the study-level raw data are generated as

$$Y_{Ci1}, \ldots, Y_{Cin_{Ci}} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \qquad Y_{Ti1}, \ldots, Y_{Tin_{Ti}} \overset{\text{iid}}{\sim} \mathcal{N}(\Delta_i, 1).$$

Let

$$\bar{Y}_{Ci} = \frac{1}{n_{Ci}} \sum_{j=1}^{n_{Ci}} Y_{Cij}, \qquad \bar{Y}_{Ti} = \frac{1}{n_{Ti}} \sum_{j=1}^{n_{Ti}} Y_{Tij},$$

and let the pooled standard deviation be

$$S_{pi} = \left[\frac{(n_{Ti} - 1) S_{Ti}^2 + (n_{Ci} - 1) S_{Ci}^2}{N_i - 2} \right]^{1/2}.$$

The simulated Cohen's d and the Hedges correction factor are

$$d_i = \frac{\bar{Y}_{Ti} - \bar{Y}_{Ci}}{S_{pi}}, \; J_i = 1 - \frac{3}{4(N_i - 2) - 1}.$$

The observed study estimate is therefore

$$Y_i = J_i d_i,$$

and the within-study sampling variance is

$$V_i = \left(\frac{N_i}{n_{Ti} n_{Ci}} + \frac{d_i^2}{2N_i} \right) J_i^2.$$

The natural significance of each study is assessed with the two-sample Student t test on the simulated raw observations, exactly as a primary study would report it. The associated test statistic and two-sided p-value are

$$T_i = \frac{\bar{Y}_{Ti} - \bar{Y}_{Ci}}{S_{pi}\sqrt{\frac{1}{n_{Ti}} + \frac{1}{n_{Ci}}}} = d_i \sqrt{\frac{n_{Ti} n_{Ci}}{N_i}}, \qquad P_i = 2\Big(1 - F_{t,\mathrm{df}_i}(|T_i|)\Big),$$

where $F_{t,\mathrm{df}}$ is the cumulative distribution function of the Student t distribution with df degrees of freedom. In an earlier version of the algorithm the p-value was instead derived from the standardised effect size as $P_i = 2(1 - F_{t,\mathrm{df}_i}(|Y_i/\sqrt{V_i}|))$; the two definitions give numerically very similar values and leave all results unchanged, as verified in Supplementary Material S5.

The simulation partitions the conducted studies into three sets:

$$A = \{i : P_i \le 0.05\}, \qquad B = \{i : 0.05 < P_i \le 0.30\}, \qquad C = \{i : P_i > 0.30\}.$$

Set A contains the originally significant studies, B contains the non-significant studies eligible for p-hacking, and C contains the remaining non-significant studies. Let

$$a = |A|, \qquad b = |B|, \qquad c = |C|, \qquad M = a + b + c.$$

Note that we retain only studies with finite Y_i, finite V_i, and $V_i > 0$; all formulas below are conditional on the retained conducted literature.

Let

$$q \in \{0.65, 0.95\}$$

denote the target proportion of published studies that must be significant (q ranges from 0.30 to 0.95 in the parameter sweep). We have

$$s^\star = \mathrm{round}(qk), \qquad \ell^\star = k - s^\star.$$

so that $s^\star$ is the target number of significant published studies and $\ell^\star$ is the target number of non-significant published studies.

If $a \ge s^\star$, then no p-hacking is needed. If $a < s^\star$, the number of additional significant studies that must be created by p-hacking is $h = s^\star - a$.

In all cases define

$$s = \min(a, s^\star).$$

which is the number of originally significant studies that are actually published.

A generation attempt is accepted only if the algorithm can construct the published sample and the resulting sample size falls within a tolerance window of the target. Let $\delta = \max(3, \text{round}(0.15\,k))$. The exact success event is

$$\mathcal{E} = \{M \geq K\} \cap \{b \geq h\} \cap \{M - a - h \geq \ell^\star\} \cap \{|k_{\text{actual}} - k| \leq \delta\},$$

where $k_{\text{actual}} = s + h + \ell^\star_{\text{actual}}$ is the realised number of published studies. The first condition is the code's requirement that at least K valid studies remain after filtering ($K = 2k$ in the main analysis). The second condition ensures that there are enough eligible studies for p-hacking. The third condition ensures that after removing the hacked studies from the non-significant pool there are still enough remaining non-significant studies to fill the $\ell^\star$ non-significant publication slots. The fourth condition is the code's acceptance tolerance, which in practice is rarely binding ($\delta = 3$ for $k = 23$).

All expectations corresponding to the actual simulation are therefore conditional on $\mathcal{E}$.

Exact construction of the published sample

Conditional on the realised conducted literature and on $\mathcal{E}$, we construct the published literature as follows.

If $s > 0$, the set of originally significant published studies, denoted S, is drawn uniformly without replacement from A so that $|S| = s$. Hence

$$\Pr(S = \mathcal{S} \mid A, \mathcal{E}) = \binom{a}{s}^{-1}$$

for every $\mathcal{S} \subseteq A$ such that $|\mathcal{S}| = s$.

If $h > 0$, eligible p-hackable studies in B are ordered by increasing p-value:

$$P^B_{(1)} \leq P^B_{(2)} \leq \cdots \leq P^B_{(b)}.$$

The hacked set is then chosen deterministically as the h studies closest to significance:

$$H = \{i_{(1)}, \ldots, i_{(h)}\}.$$

Let

$$r = \left\lceil \frac{h}{2} \right\rceil.$$

The first r hacked studies, namely $i_{(1)}, \ldots, i_{(r)}$, undergo "mild" p-hacking, while the remaining $h-r$ hacked studies, namely $i_{(r+1)}, \ldots, i_{(h)}$, undergo "aggressive" p-hacking.

For the mild subgroup, the reported effect size is unchanged:

$$Y^{\star}_{i_{(j)}} = Y_{i_{(j)}}, \qquad j = 1, \dots, r,$$

and only the reported p-value is replaced by an independent draw

$$P^{\star}_{i_{(j)}} \sim \mathrm{Unif}(0.01, 0.049).$$

For the aggressive subgroup, we generate

$$U_{i_{(j)}} \overset{\text{iid}}{\sim} \mathrm{Unif}(0.01, 0.049), \qquad j = r+1, \dots, h,$$

then define

$$T^{\star}_{i_{(j)}} = F^{-1}_{t,\mathrm{df}_{i_{(j)}}} \left(1 - \frac{U_{i_{(j)}}}{2}\right).$$

The observed effect size becomes

$$Y^{\star}_{i_{(j)}} = T^{\star}_{i_{(j)}} \sqrt{V_{i_{(j)}}}.$$

The reported p-value is set equal to

$$P^{\star}_{i_{(j)}} = U_{i_{(j)}}.$$

Note the within-study variance is not altered, so for all hacked studies $V^{\star}_{i_{(j)}} = V_{i_{(j)}}$.

The remaining non-significant studies are

$$R = (B \cup C) \setminus H.$$

The published non-significant set N is sampled uniformly without replacement from R so that $|N| = \ell^{\star}$. Hence

$$\Pr(N = \mathcal{N} \mid R, \mathcal{E}) = \binom{|R|}{\ell^{\star}}^{-1}$$

for every $\mathcal{N} \subseteq R$ such that $|\mathcal{N}| = \ell^{\star}$.

The final published set is

$$\mathcal{P} = S \cup H \cup N, \qquad |\mathcal{P}| = k.$$

Expected effect among originally significant published studies

Because S is a simple random sample without replacement from A, the sample average of the originally significant published studies is unbiased for the finite-population average over A. Therefore, conditional on the realised conducted literature,

$$\mathbb{E}\left(\frac{1}{s}\sum_{i \in S} Y_i \,\middle|\, A, \mathcal{E}\right) = \frac{1}{a}\sum_{i \in A} Y_i, \qquad a > 0.$$

Equivalently, the exact expected effect among originally significant published studies is

$$\mu_S^{\text{pub}}(D) = \frac{1}{a}\sum_{i\in A} Y_i, \qquad a > 0,$$

where D denotes the realised conducted literature.

Passing from the realised literature to the simulation distribution, the exact expected effect among originally significant published studies is

$$\mu_S^{\text{pub}} = \mathbb{E}\left(\frac{1}{|A|}\sum_{i\in A} Y_i \,\middle|\, |A| > 0, \mathcal{E}\right).$$

By exchangeability of the K conducted studies, this can also be written as

$$\mu_S^{\text{pub}} = \mathbb{E}\left(Y_1 \mid P_1 \le 0.05, \mathcal{E}\right).$$

Expected reported effect among hacked studies

Let

$$\bar{Y}_H^\star = \frac{1}{h}\sum_{i\in H} Y_i^\star$$

denote the average reported effect among hacked studies, defined only when $h > 0$. Conditional on the realised conducted literature D and on the success event $\mathcal{E}$, the hacked set H is deterministic, and randomness arises only from the uniforms used in the severe hacking step. Hence

$$\mathbb{E}(\bar{Y}_H^\star \mid D, \mathcal{E}, h > 0) = \frac{1}{h}\left[\sum_{j=1}^{r} Y_{i_{(j)}} + \sum_{j=r+1}^{h} \mathbb{E}\left(Y_{i_{(j)}}^\star \,\middle|\, D, \mathcal{E}\right)\right].$$

For $j = r+1, \ldots, h$,

$$Y_{i_{(j)}}^\star = \sqrt{V_{i_{(j)}}}\, F_{t,\mathrm{df}_{i_{(j)}}}^{-1}\left(1 - \frac{U_{i_{(j)}}}{2}\right), \qquad U_{i_{(j)}} \sim \mathrm{Unif}(0.01, 0.049),$$

so

$$\mathbb{E}\left(Y_{i_{(j)}}^\star \,\middle|\, D, \mathcal{E}\right) = \sqrt{V_{i_{(j)}}}\, m(\mathrm{df}_{i_{(j)}}),$$

where

$$m(d) = \frac{1}{0.039}\int_{0.01}^{0.049} F_{t,d}^{-1}\left(1 - \frac{u}{2}\right) du.$$

Therefore the exact conditional expectation among hacked studies is

$$\mu_H(D) := \mathbb{E}(\bar{Y}_H^\star \mid D, \mathcal{E}, h > 0) = \frac{1}{h}\left[\sum_{j=1}^{r} Y_{i_{(j)}} + \sum_{j=r+1}^{h} \sqrt{V_{i_{(j)}}}\, m(\mathrm{df}_{i_{(j)}})\right].$$

Averaging over the simulation law yields the exact unconditional expectation

$$\mu_H = \mathbb{E}\left(\mu_H(D) \,|\, h > 0, \mathcal{E}\right).$$

Equivalently,

$$\mu_H = \mathbb{E}\left[\frac{1}{h}\left(\sum_{j=1}^{\lceil h/2 \rceil} Y_{i_{(j)}} + \sum_{j=\lceil h/2 \rceil + 1}^{h} \sqrt{V_{i_{(j)}}}\, m(\mathrm{df}_{i_{(j)}})\right) \middle| h > 0, \mathcal{E}\right].$$

Expected reported effect among all significant published studies

The set of significant published studies is

$$G = S \cup H, \qquad |G| = s + h = s^\star.$$

Its average reported effect is

$$\bar{Y}_G^\star = \frac{1}{s^\star}\left(\sum_{i \in S} Y_i + \sum_{i \in H} Y_i^\star\right).$$

Conditional on D and $\mathcal{E}$,

$$\mathbb{E}(\bar{Y}_G^\star \mid D, \mathcal{E}) = \frac{1}{s^\star}\left[\frac{s}{a}\sum_{i \in A} Y_i + \sum_{j=1}^{r} Y_{i_{(j)}} + \sum_{j=r+1}^{h} \sqrt{V_{i_{(j)}}}\, m(\mathrm{df}_{i_{(j)}})\right].$$

Exact expectation and bias of the CMA estimator

In the simulation, an intercept-only random-effects meta-analysis model is fitted to the published sample by REML. Let

$$\mathbf{Y}^\star = (Y_1^\star, \ldots, Y_k^\star)^\top, \qquad \mathbf{V}^\star = (V_1^\star, \ldots, V_k^\star)^\top$$

denote the published effect sizes and within-study variances after the publication and p-hacking algorithm has been applied. The published-study ordering is immaterial. Define

$$\hat{\tau}^2_{\mathrm{REML}}(\mathbf{y}, \mathbf{v})$$

to be the REML estimate of the between-study variance in the intercept-only model returned by the `rma` fit, and define

$$\psi_{\mathrm{REML}}(\mathbf{y}, \mathbf{v}) = \frac{\sum_{j=1}^{k} \dfrac{y_j}{v_j + \hat{\tau}^2_{\mathrm{REML}}(\mathbf{y}, \mathbf{v})}}{\sum_{j=1}^{k} \dfrac{1}{v_j + \hat{\tau}^2_{\mathrm{REML}}(\mathbf{y}, \mathbf{v})}}.$$

Then the CMA point estimator is exactly

$$\hat{\mu}_{\mathrm{CMA}} = \psi_{\mathrm{REML}}(\mathbf{Y}^\star, \mathbf{V}^\star).$$

Conditional on the realised conducted literature D and on $\mathcal{E}$, the only remaining randomness comes from the uniform subset S of originally significant studies, the uniform subset N of non-significant published studies, and the uniforms used in the severe p-hacking step. Hence the exact conditional expectation of the CMA estimator is

$$\mathbb{E}(\hat{\mu}_{\mathrm{CMA}} \mid D, \mathcal{E}) = \frac{1}{\binom{a}{s}\binom{M-a-h}{\ell^\star}} \sum_{\substack{\mathcal{S}\subseteq A \\ |\mathcal{S}|=s}} \sum_{\substack{\mathcal{N}\subseteq (B\cup C)\setminus H \\ |\mathcal{N}|=\ell^\star}} \left[\frac{1}{0.039^{\,h-r}} \int_{[0.01,0.049]^{h-r}} \psi_{\mathrm{REML}}\big(\mathbf{y}^\star, \mathbf{v}^\star\big)\, d\mathbf{u} \right],$$

where

$$\psi_{\mathrm{REML}}\big(\mathbf{y}^\star, \mathbf{v}^\star\big) = \psi_{\mathrm{REML}}\big(\mathbf{y}^\star(D, \mathcal{S}, \mathcal{N}, \mathbf{u}), \mathbf{v}^\star(D, \mathcal{S}, \mathcal{N})\big),$$

$\mathbf{y}^\star(D, \mathcal{S}, \mathcal{N}, \mathbf{u})$ is the k-vector of published effect sizes formed by concatenating

$$\{Y_i : i \in \mathcal{S}\},\ \{Y_{i_{(j)}} : j = 1, \ldots, r\},\ \left\{ \sqrt{V_{i_{(j)}}}\, F^{-1}_{t,\mathrm{df}_{i_{(j)}}} \left(1 - \frac{u_j}{2}\right) : j = r+1, \ldots, h \right\},\ \{Y_i : i \in \mathcal{N}\},$$

and $\mathbf{v}^\star(D, \mathcal{S}, \mathcal{N})$ is the matching vector of within-study variances, which is simply the corresponding collection of original V_i's.

The exact unconditional expectation of the CMA estimator under the simulation is therefore

$$\mathbb{E}(\hat{\mu}_{\mathrm{CMA}} \mid \mathcal{E}) = \mathbb{E}\left[\mathbb{E}(\hat{\mu}_{\mathrm{CMA}} \mid D, \mathcal{E}) \,|\, \mathcal{E}\right].$$

Accordingly, the exact bias of the CMA estimator induced by the simulation algorithm is

$$\mathrm{Bias}^{\mathrm{REML}}_{\mathrm{CMA}} = \mathbb{E}(\hat{\mu}_{\mathrm{CMA}} \mid \mathcal{E}) - \mu.$$

This expression is exact for finite sample representation as defined by the algorithm. It is not reducible to a simpler elementary closed form without some approximations and inexact simplifications.

Supplementary Material S3: Closed-form approximation for CMA bias

We derive a closed-form approximation for CMA bias under selective publication and p-hacking, using the truncated normal distribution. This avoids the combinatorial complexity of the exact expressions in S2 while closely matching the simulation results (Table 2).

Setup

We use the notation of S2 throughout. Each conducted study i has true effect $\Delta_i \sim \mathcal{N}(\mu, \tau^2)$, sample size N_i, observed effect $Y_i = J_i d_i$ (Hedges' g), and within-study variance $V_i \approx (N_i / n_{Ti} n_{Ci})\, J_i^2$. A study is significant when $|Y_i| > Y_{\text{crit},i}$, where $Y_{\text{crit},i} = t_{0.975,\, N_i - 2}\sqrt{V_i}$. As in S2, $s^\star = \text{round}(qk)$ studies in the published sample must be significant and $\ell^\star = k - s^\star$ are non-significant.

Assumptions

We make two simplifications. First, conditional on (Δ_i, N_i):

$$Y_i \mid \Delta_i, N_i \;\dot{\sim}\; \mathcal{N}(J_i \Delta_i,\; V_i).$$

Second, we approximate the REML-weighted estimator $\hat{\mu}_{\text{CMA}} = \psi_{\text{REML}}(\mathbf{Y}^\star, \mathbf{V}^\star)$ (S2) by a simple average of published effects. This is reasonable because the REML weights $1/(V_i + \hat{\tau}^2)$ become approximately uniform as $\hat{\tau}^2$ grows.

Truncated normal expectations

For a study with $\Delta_i = \delta$, $N_i = n$, let $\sigma_i = \sqrt{V_i}$, $\bar{Y}_i = J\delta$, and define

$$\alpha_u = \frac{Y_{\text{crit}} - \bar{Y}_i}{\sigma_i}, \qquad \alpha_\ell = \frac{-Y_{\text{crit}} - \bar{Y}_i}{\sigma_i},$$

where ϕ and Φ are the standard normal density and CDF. The expected effect among significant studies (right tail, which dominates for positive effects) is

$$\mathbb{E}[Y_i \mid Y_i > Y_{\text{crit}},\, \delta, n] \;=\; \bar{Y}_i + \sigma_i \cdot \frac{\phi(\alpha_u)}{1 - \Phi(\alpha_u)},$$

with probability of significance $\pi(\delta, n) = 1 - \Phi(\alpha_u) + \Phi(\alpha_\ell)$. For non-significant studies (doubly-truncated normal):

$$\mathbb{E}[Y_i \mid |Y_i| < Y_{\text{crit}},\, \delta, n] \;=\; \bar{Y}_i + \sigma_i \cdot \frac{\phi(\alpha_\ell) - \phi(\alpha_u)}{\Phi(\alpha_u) - \Phi(\alpha_\ell)}.$$

Integration over the study population

Since Δ_i and N_i vary across studies, we integrate over $\delta \sim \mathcal{N}(\mu, \tau^2)$ and average over sample-size strata $\{(n_j, w_j)\}$ (midpoints: $n = 25, 50, 85$ with weights $0.25, 0.50, 0.25$):

$$\mathbb{E}[Y \mid \text{sig}] = \frac{\displaystyle\int \sum_j w_j \, \mathbb{E}[Y \mid Y > Y_{\text{crit}},\, \delta, n_j] \; \pi(\delta, n_j) \; f(\delta) \, d\delta}{\displaystyle\int \sum_j w_j \; \pi(\delta, n_j) \; f(\delta) \, d\delta},$$

where $f(\delta)$ is the $\mathcal{N}(\mu, \tau^2)$ density. The non-significant expectation $\mathbb{E}[Y \mid \text{non-sig}]$ is analogous. The marginal significance probability is $\bar{\pi} = \int \sum_j w_j \, \pi(\delta, n_j) \, f(\delta) \, d\delta$.

P-hacking components

When $K\bar{\pi} < s^\star$, the shortfall $h = s^\star - K\bar{\pi}$ studies require p-hacking (S2). The first $r = \lceil h/2 \rceil$ undergo mild hacking (effect Y_i unchanged), the remaining $h - r$ undergo aggressive hacking. Their expected effects are:

Mild: these studies have $p \in (0.05, 0.30]$; their original effects sit between the $p = 0.30$ and $p = 0.05$ boundaries, so the expected original effect is approximated as

$$\mathbb{E}[Y \mid \text{mild}] \;\approx\; \tfrac{1}{2}(Y_{\text{low}} + Y_{\text{crit}}),$$

where $Y_{\text{low}} = t_{0.85,\,\bar{N}-2}\,\bar{\sigma}$ is the $p = 0.30$ boundary.

Aggressive: following S2, $Y_i^\star = T_i^\star \sqrt{V_i}$ with $T_i^\star = F^{-1}_{t,\mathrm{df}_i}(1 - U_i/2)$, $U_i \sim \text{Unif}(0.01, 0.049)$, giving

$$\mathbb{E}[Y^\star \mid \text{aggressive}] \;=\; \bar{\sigma} \cdot m(\bar{d}), \qquad m(d) = \frac{1}{0.039} \int_{0.01}^{0.049} F^{-1}_{t,d}\big(1 - \tfrac{u}{2}\big)\, du \quad \text{(S2)}.$$

Combined approximation

With $s = \min(K\bar{\pi},\, s^\star)$ originally significant published studies:

$$\mathbb{E}[\hat{\mu}_{\text{CMA}}] \;\approx\; \frac{1}{k}\Big[\, s\,\mathbb{E}[Y \mid \text{sig}] + r\,\mathbb{E}[Y \mid \text{mild}] + (h-r)\,\mathbb{E}[Y^\star \mid \text{aggressive}] + \ell^\star\,\mathbb{E}[Y \mid \text{non-sig}] \Big],$$

$$\text{Bias}^{\text{approx}}_{\text{CMA}} = \mathbb{E}[\hat{\mu}_{\text{CMA}}] - \mu.$$

When $h = 0$ (no p-hacking needed), this simplifies to $\mathbb{E}[\hat{\mu}_{\text{CMA}}] \approx (s^\star/k)\,\mathbb{E}[Y \mid \text{sig}] + (\ell^\star/k)\,\mathbb{E}[Y \mid \text{non-sig}]$.

These expressions allow researchers to obtain a quick estimate of expected CMA bias for their own field-specific parameters—effect size magnitude, heterogeneity, sample sizes, and target significance rate—without running the full simulation. An accompanying R script (`S3.R`) is provided in the repository.

Interpretation

The bias is driven by the ratio of the significance threshold Y_{crit} to the true effect. When $Y_{\text{crit}} \gg \mu$ (small effects, small samples), only studies with large positive sampling errors pass the filter, and the uplift $\phi(\alpha_u)/(1-\Phi(\alpha_u))$ is large. When μ is large enough that most studies are naturally significant, the truncation point falls into the left tail and the uplift vanishes.

Increasing τ^2 widens the distribution of Δ_i, so more studies achieve significance through genuinely large effects rather than sampling luck, attenuating the selection bias. This matches the simulation finding that RDMA's advantage shrinks with heterogeneity.

CI coverage and error bounds

Under selective publication, the CMA estimate is centred on $\mu + \text{Bias}$ while the CI is constructed as if unbiased. With half-width $w = z_{0.975} \cdot \text{SE}$:

$$\text{Coverage} \;\approx\; \Phi\left(\frac{w - \text{Bias}}{\text{SE}}\right) - \Phi\left(\frac{-w - \text{Bias}}{\text{SE}}\right).$$

When Bias $\gg w$, coverage collapses, consistent with 21% coverage for small effects at a 95% significance rate. The MAE decomposes as $|\text{Bias}| + \mathbb{E}[|\varepsilon|]$ where $\varepsilon \sim O(\sqrt{(\bar{V} + \tau^2)/k})$. The bias is a structural property of the publication filter and does not diminish with k; only the random component shrinks. RDMA eliminates the bias term entirely.

Validation

Table 2 compares the approximation with simulation (Table 1, 10,000 iterations). Mean absolute difference: 0.012; maximum: 0.041. The largest gaps occur for small effects at a 95% significance rate with low heterogeneity, where p-hacking is heaviest and the midpoint approximation for borderline studies is coarsest.

Table 2: Approximation vs. simulation for $\mathbb{E}[\hat{\mu}_{\text{CMA}}]$

True ES	**Het.**	**Sig. rate**	**Approx.**	**Sim.**	**Diff.**
0.2	Low	65%	0.394	0.37	+0.024
0.2	Low	95%	0.521	0.48	+0.041
0.2	Medium	65%	0.331	0.33	+0.001
0.2	Medium	95%	0.478	0.44	+0.038
0.2	High	65%	0.302	0.30	+0.002
0.2	High	95%	0.433	0.41	+0.023
0.5	Low	65%	0.628	0.63	−0.002
0.5	Low	95%	0.824	0.80	+0.024
0.5	Medium	65%	0.581	0.58	+0.001
0.5	Medium	95%	0.807	0.80	+0.007
0.5	High	65%	0.546	0.55	−0.004
0.5	High	95%	0.767	0.77	−0.003
0.8	Low	95%	1.073	1.07	+0.003
0.8	Medium	95%	1.096	1.10	−0.004
0.8	High	95%	1.071	1.08	−0.009

Supplementary Material S4: Sensitivity of results to p-hacking assumptions

Two features of the p-hacking procedure are not pinned down by empirical evidence: the p-value window within which a non-significant study is eligible for p-hacking, and the split between threshold-shifting (mild) and estimate-inflating (aggressive) manipulation. We varied each in turn, holding all other parameters at the empirical ecological values, at 2,000 iterations per cell. **Eligibility window.** We repeated the ecological case-study scenarios with the window set to $0.05 < p \leq w$ for $w \in \{0.20, 0.30, 0.40, 0.50\}$. Widening the window admits studies further from significance into the eligible pool, but because the algorithm always p-hacks the studies closest to significance first, the composition of the published literature barely changes: RDMA's advantage moved by at most 0.014 across the entire range of w, in every scenario. **Mild versus aggressive split.** We varied the proportion of manipulated studies subjected to threshold-shifting rather than estimate-inflating manipulation across $\{0, 0.25, 0.50, 0.75, 1.00\}$, that is, from all manipulation inflating the estimate to all manipulation only shifting significance. Across all 75 combinations of split and scenario, RDMA remained more accurate than CMA (Figure 6). The direction of the conclusion therefore does not depend on which mode of p-hacking is assumed. As expected, the magnitude of RDMA's advantage is larger when manipulation inflates the effect size, because CMA is then more distorted, and smaller when manipulation only shifts the p-value across the threshold while leaving the effect size unchanged. The latter extreme, in which all manipulation is threshold-shifting, is therefore the one least favourable to RDMA; even there, RDMA remained more accurate than CMA. Our main analysis uses an intermediate 50% split, so it neither maximises nor minimises the reported advantage.

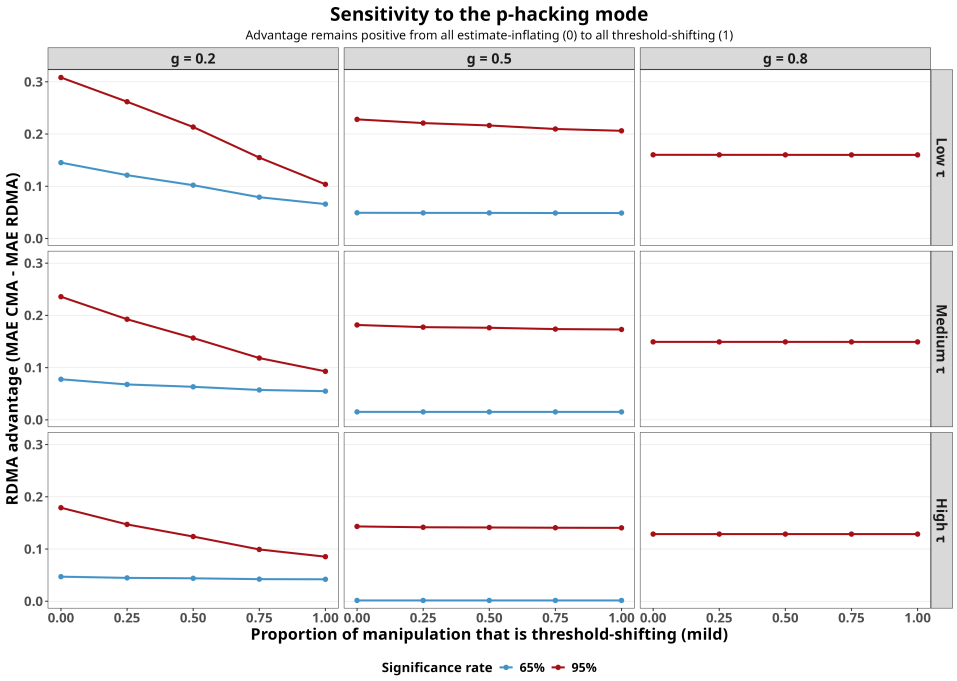


Figure 6: RDMA advantage (mean absolute error of CMA minus that of RDMA) as a function of the proportion of manipulation that is threshold-shifting, from 0 (all estimate-inflating) to 1 (all threshold-shifting), for each effect size (columns) and heterogeneity level (rows), at 65% and 95% significance rates. RDMA is more accurate throughout.

Supplementary Material S5: Definition of each study's natural significance

Each study's natural significance can be assessed either from a two sample Student t test on the simulated raw observations, exactly as a primary study would report it, or from the standardised effect size and its sampling variance as $P_i = 2(1 - F_{t,\mathrm{df}_i}(|Y_i/\sqrt{V_i}|))$. The submitted version of the code used the second; following Reviewer 1, the revised version uses the first, which is the more faithful representation of how a primary study reports a result and of the quantity on which publication selection and p-hacking act. We reran all fifteen headline scenarios under both definitions. The two give numerically very similar results: across the fifteen scenarios the mean absolute error of both methods differed by at most 0.02 (mean 0.004), the accuracy of heterogeneity estimation by at most 0.009, confidence interval coverage by at most 4.4 percentage points, and the mean percentage of studies requiring p-hacking by at most 2.8 percentage points (for example, from 35.1% to 32.4% for small effects with medium heterogeneity at a 95% significance rate). The raw-data definition finds marginally more studies naturally significant, so slightly less p-hacking is required to reach each target significance rate. The sign of RDMA's advantage is unchanged in every scenario. Because the two-sample t test consumes no random numbers, the two definitions share an identical simulation stream up to the point at which a study is classified as significant, so these differences reflect the change of definition alone and not run-to-run variation. All results reported in the main text use the raw-data definition.

Supplementary Material S6: Full parameter-sweep results

We crossed the publication rate (20%, 50%, 70%) with the significance rate in the published literature (30% to 95% in steps of 5%), for each of the three effect sizes and three heterogeneity levels, at 2,000 iterations per cell (297 valid cells in total; cells requiring a published literature more than five percentage points less significant than the natural rate are omitted). Figure 5 in the main text summarises RDMA's advantage across this landscape.

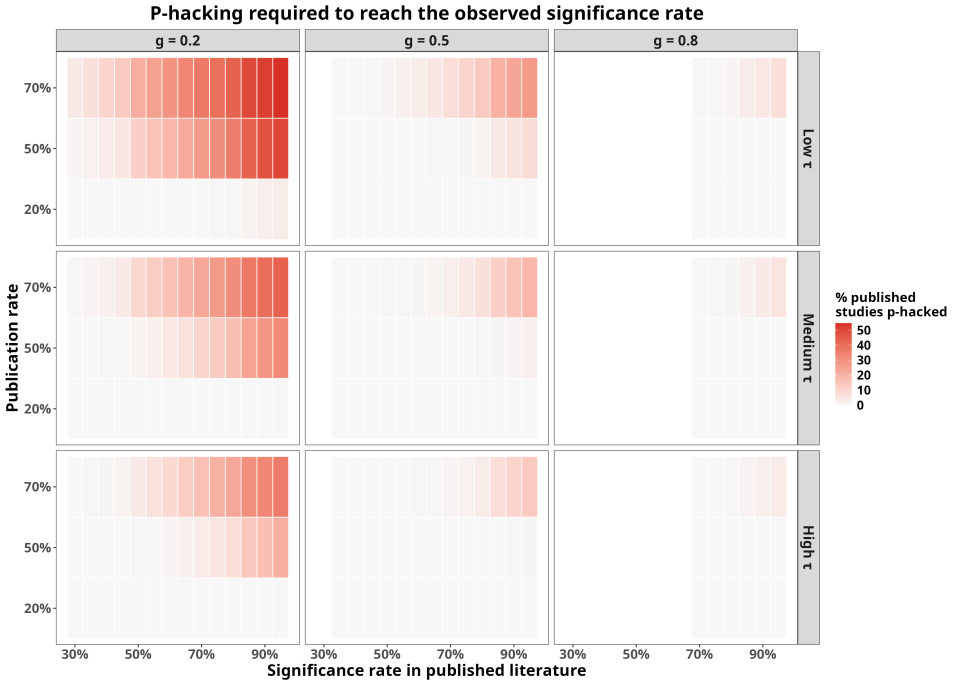


Figure 7: Mean percentage of published studies requiring p-hacking to reach the target significance rate, across the publication rate and the significance rate in the published literature, for each effect size (columns) and heterogeneity level (rows).

Figure 7 shows the companion surface: the mean percentage of published studies that must have been p-hacked to reach each significance rate. P-hacking is negligible at low significance rates and at large effect sizes, and rises steeply towards small effects, high significance rates

and high publication rates, the corner in which the unpublished pool is smallest relative to the number of significant studies required. Every one of the 297 cells could be constructed; only the single most extreme combination ($g = 0.2$, low heterogeneity, 95% significance rate, 70% publication rate) fell short of the full 2,000 iterations, completing 1,371 of them. The complete cell-level results are available in the data repository (NEW_sweep_full.zip, NEW_sweep_heat_effect.csv, NEW_sweep_feasibility.csv).