A Statistical Case for Qualified Scientific Optimism

Erik van Zwet* Andrew Gelman† Witold Więcek‡

22 Sep 2026

Abstract

We create an open-source database from empirical research in medicine, economics, psychology, psychotherapy, ecology, evolution, political science, education, and more, and make it available at `https://github.com/wwiecek/BEAR/`. Each result is represented by a z-value: an estimated effect divided by its standard error.

We then provide a modeling framework that allows us to move away from the usual attempts to classify effects as exactly zero or not, and instead focus on identifying the direction of effects and their replicability.

After adjusting for publication bias, we find two patterns that hold across all datasets and their various subsets: (1) statistical significance alone does not imply a high probability of successful replication, but (2) it does imply a high probability that the direction of the observed effect was correct.

We conclude that replication rate is a poor measure of the quality of a field of research and argue that an attitude of global scientific pessimism based on low replication rates is misguided. Instead, we favor a more optimistic view of scientific research based on the ability to identify the direction of effects. This optimism is not naive as we recognize the many challenges science continues to face such as questionable research practices, paper mills, fake AI papers, and such.

1. Studying the quality of empirical science

The replication crisis has made us aware that there's a wide range of quality of studies in the life and behavioral sciences, even considering subsets such as particular research areas and papers published in top journals. This makes sense; if we roughly measure the quality of a study by its signal-to-noise ratio (the true effect size divided by its standard error), then we have some control over the denominator (by taking more or better measurements) but less control over the numerator. It's the nature of research to be uncertain and it's the duty of researchers to pursue all avenues, studying effects which might turn out to be null or highly context-dependent.

A key turning point in the replication crisis was the recognition and refutation of the false intuition that you can tell whether a study is good or not (tell whether a result is strong or not) using a statistical significance threshold. Beyond the problems with any sharp threshold, even seemingly clear distinctions such as, for example, $p < 0.01$ compared to $p > 0.2$ do not discriminate well. This is not a problem that is specific to p-values. Rather, the problem is that a single study rarely—if ever—provides definitive evidence. Thus, similar problems arise not just with p-values but also with Bayes factors and other measures of evidence.

But this should not be taken as a negative verdict on statistically-based science as a whole. To study this, we have created an extensive open-source database of empirical research results and focus here on 16 publicly available corpora in diverse fields. Our optimistic finding is that, in these corpora, most statistically significant results seem to be correct in sign. Our remaining pessimism stems from clear signs of selection in analysis and reporting of results in some of the corpora, which points toward the need for real replication—not in the hope of confirming significance but to understand effects, and to minimize selection in all phases of the process.

*Department of Biomedical Data Sciences, Leiden University Medical Center.

†Department of Statistics and Department of Political Science, Columbia University, New York.

‡Development Innovation Lab, University of Chicago.

It is well known to most that you shouldn't rely on a single study and that replication and, ultimately, meta-analysis are the way to go. But then many people are still dismayed when results "fail to replicate." One point of the present paper is that replication rates in a research area can be low even when most studies are finding true effects. Hence low replication rates should not be taken as implying that most of science is wrong.

1.1. Going beyond the characterization of scientific claims as true or false

In an influential paper, Ioannidis (2005) viewed empirical research as "a 2×2 table in which research findings are compared against the gold standard of true relationships in a scientific field."

This 2×2 model is implicitly accepted in the meta-scientific literature, which tends to focus on statistical significance and replication rates (Simmons, Nelson, and Simonsohn 2011; Open Science Collaboration 2015; Camerer et al. 2016). In particular, low replication rates are taken by many as evidence of published empirical research being full of null claims and hence not credible as a whole, a position with political influence; see for example Garisto (2025).

We understand that the 2×2 model was always intended as an instructive simplification, introduced for the purpose of explaining that in an area of research where the power is low and there are few true relationships, many statistically significant results may not be true. However, far-reaching conclusions from this model have been made about the actual state of scientific research, in particular the claim that more than half of all published research findings are false.

We have two main criticisms of the 2×2 model. First, we disagree with the notion that the primary goal of scientific study is to separate "no relationships" (or "nulls") from "true relationships." Effects of interest are rarely exactly zero, and even if they are, they will no longer be zero if one takes leakage, expectation effects, dropout, etc. into account. But apart from that, with finite data, it is simply not possible to distinguish effects that are zero from those that are small.

Our second criticism is that the 2×2 model is not a realistic representation of empirical research. The key feature of the 2×2 model is that the probability that a study from some area of research reaches statistical significance only takes two values: 5% for the null-studies and some other, much larger probability (like 80%) for the non-null studies. That is simply not possible. Researchers would have to know the true effect (if one exists) to be able to design their studies to have either 5% or 80% probability of reaching significance. In reality, researchers do not know how large the true effect is, and, moreover, the design of studies is also based on external considerations such as time, money and the availability of participants.

Here we make an effort to build more realistic statistical models of scientific research. We have compiled a database of corpora from medicine, economics, psychology, psychotherapy, ecology, evolution, political science, education, and more.

Concretely, we make two contributions. First, we make a large number of results from empirical research easily accessible in a new database, Benchmarks of Empirical Accuracy in Research (BEAR). We do not contribute any new data, but we process the available data to make them ready for meta-scientific research. We provide clear documentation and all data are available under open source license at `https://github.com/wwiecek/BEAR/`. We will continue to add new datasets as they become available.

Second, we extend the meta-scientific model of Stephens (2017), van Zwet, Schwab, and Senn (2021), and van Zwet and Gelman (2022) to account for truncation and publication bias. This model allows us to study measures of research quality that are insensitive to whether effects are exactly zero or merely very small, such as the probability of correctly identifying the direction of an effect and the probability of replication (Greenland 2017; Gelman and Carlin 2014; Gelman and Tuerlinckx 2000).

1.2. Summary of findings

We conclude from our analyses that statistical significance usually implies a high probability that the direction of the observed effect is correct, but that it does not imply a high probability that a hypothetical exact replication study would reach statistical significance. In other words, it is expected that true findings (in the sense of correctly identifying the direction of an effect) will often not replicate. Similar conclusions were reached by Bak-Coleman et al. (2022) and Neves, Tan, and Amaral (2022) on the basis of simulations and data from replication studies.

The scientific process is far from perfect, and questionable research practices, paper mills, and outright fraud are serious concerns (Aquarius et al. 2025; Matusz, Abalkina, and Bishop 2025). But the mere fact that many studies have low probability of reaching statistical significance and therefore fail to replicate does not imply that most published research findings are false.

The challenge is internalizing the juxtaposition of two ideas. On one hand, there really is a replication crisis: substantive and statistical theory, empirical evidence, and sociological reasoning all point toward the conclusion that traditional procedures for scientific quality control (peer review, causal identification, and statistical significance) are not enough to ensure the reliability of research findings. On the other hand, the statistical evidence implies that in many fields, published results that reach $p < 0.05$ generally fall in the correct direction.

2. Data

We have compiled a large number of empirical results from a variety of sources in a new database which we call Benchmarks of Empirical Accuracy in Research (BEAR). BEAR currently has 26 dataset entries, but for this paper we use 16 of those, comprising about 180,000 studies. Characteristics of the presented datasets are given in Table 1 and in Appendix B. We describe our processing steps in Appendix C.

We distinguish three groups among our datasets. The first group contains eight curated sets. These include meta-analytic or systematic-review collections in intelligence research (Nuijten et al. 2020), economics (Askarov et al. 2023), psychotherapy (Cuijpers, Harrer, Miguel, et al. 2025), political science (Arel-Bundock et al. 2022), ecology and evolution (Costello and Fox 2022; Yang, van Zwet, et al. 2024), psychology (Rodriguez and Williams 2022), and medicine through the Cochrane Database of Systematic Reviews (Schwab 2024; Cochrane Collaboration 2025). We also include education intervention results from the What Works Clearinghouse (Institute of Education Sciences 2025). Where the source is meta-analytic, BEAR treats the rows as collections of individual studies, while retaining meta-analysis identifiers for other uses.

The second group consists of six corpora that can be thought of as representative samples of studies or results within research domains: SCORE dataset of scientific claims in the social and behavioral sciences (Tyner et al. 2026), main results from top economics papers (Lang 2025), text-mined t-tests in cognitive neuroscience (Szucs and Ioannidis 2017), clinical trial registry results from `clinicaltrials.gov` and EU CTR (National Library of Medicine 2025; European Medicines Agency 2025), and text-mined p-values (Chavalarias et al. 2016) and confidence intervals (Barnett and Wren 2019) in biomedical journals.

For comparison, we also include a third group which has the studies that that were selected for replication in two large scale replication projects: the Open Science Collaboration replications in psychology (Open Science Collaboration 2015) and the SCORE replications in the social and behavioral sciences (Tyner et al. 2026).

While we do some data processing of our own, all of the data collection and almost all of the data processing was already done by the creators of the datasets we reference. Most of these datasets

Corpus	Purpose of dataset	N	$\bar{k}$	% signif.
Curated sets				
Nuijten et al. (2020)	Meta-meta-analysis of intelligence research	1,913	1.3	53%
Askarov et al. (2023)	Economics meta-analyses and leading-journal estimates	1,913	11.2	52%
Metapsy (Harrer, Karyotaki, and Cuijpers 2022)	Meta-analytic psychotherapy trial databases	1,494	2.9	48%
Arel-Bundock et al. (2022)	Meta-analytic estimates in political science	2,252	7.4	47%
Costello and Fox (2022)	Meta-analyses of decline effects in ecology and evolution	12,927	6.8	42%
psymetadata (Rodriguez and Williams 2022)	Curated psychology meta-analysis datasets	721	11.8	36%
What Works Clearinghouse (Institute of Education Sciences 2025)	Education intervention effect sizes	1,402	5.0	34%
Cochrane Database of Systematic Reviews	Systematic reviews in health and medicine	27,513	1.3	33%
Representative domain samples				
SCORE claims (Tyner et al. 2026)	Claims from papers selected for SCORE	159	12.2	87%
Chavalarias et al. (2016)	Text-mined p-values from Medline and PubMed	50,000	1.0	87%
Barnett and Wren (2019)	Text-mined confidence intervals in health and medicine	50,000	1.0	84%
Szucs and Ioannidis (2017)	Text-mined t-tests in cognitive neuroscience	2,261	7.5	62%
Lang (2025)	Principal and abstract results from economics papers	736	5.3	58%
clinicaltrials.gov + EU CTR	Clinical trial registries in the United States and Europe	31,710	2.2	44%
Replication efforts				
SCORE replications (Tyner et al. 2026)	Matched replications of published claims	163	1.6	64%
Open Science Collaboration (2015)	Replications of psychology experiments	97	1.0	35%

Table 1: Datasets included in the paper, grouped into curated sets, representative domain samples, and replication projects. Study units are source-specific studies, trials, papers, or claims. $\bar{k}$ is the mean number of observations per study unit, and "% signif." is the percentage of absolute z-values exceeding 1.96. Barnett and Wren and Chavalarias are subsamples of 50,000 studies, with one z-value selected from each study at random; the original datasets have 416,000 and 1.9 mln values, respectively.

already include z-values or effect sizes and standard errors, but in some cases we do some conversions to z-values ourselves (for example, from p-values or 2×2 tables of counts). The details of these calculations are in the Appendix B.

Although in some cases we can extract main claims ("primary outcomes") from studies (Lang 2025; National Library of Medicine 2025), datasets typically include many z-values per study; see Table 1. In the case of Arel-Bundock et al. (2022) which aimed to collect all tests reported by economics papers, there can be as many as 50. Data are also heterogeneous in terms of the proportion of z-values that exceed 1.96 in magnitude: from 0.31 in case of the Cochrane database, up to almost 90% in SCORE and PubMed surveys.

3. Statistical model

3.1. Signal-to-noise ratios

Suppose that we have a collection of unbiased effect estimates which are normally distributed with known standard errors. The z-value is the ratio of the effect estimate to its standard error. We define the signal-to-noise ratio (SNR) as the ratio of the (unobserved) true effect to the standard error of its estimate. Our assumptions imply that the z-value is equal to the SNR plus an independent standard normal error.

The null hypothesis that there is no effect is equivalent to the hypothesis that the SNR is zero. This hypothesis is commonly rejected when the absolute value of the z-value exceeds 1.96 or, equivalently, when the associated two-sided p-value is less than 0.05. The type I error probability of this test is $\alpha = 0.05$. In our notation,

$$\Pr(|z| \geq 1.96 \mid \text{SNR} = 0) = 0.05.$$

The probability of statistical significance (PoS) depends on the true effect and the standard error through the magnitude of the SNR,

$$\text{PoS}(\text{SNR}) = \Pr(|z| \geq 1.96 \mid \text{SNR}) = \Phi(-1.96 - |\text{SNR}|) + 1 - \Phi(1.96 - |\text{SNR}|). \tag{1}$$

For example, if the SNR is 2.8 (or -2.8), then the PoS is 80%. Averaging the PoS over the distribution of the SNR in a field of research, we obtain the "assurance" which we denote by $\overline{\text{PoS}}$. It is the proportion of studies for that field that reaches the 5% level of statistical significance. We can also consider $\Pr(|\text{SNR}| \geq 2.8) = \Pr(\text{PoS}(\text{SNR}) \geq 0.8)$, which is the proportion of studies for which the PoS is 80% or more.

In the introduction, we mentioned that p-values do not discriminate well between reliable and unreliable results because they are inherently variable. This can be most easily understood by considering SNRs and z-values. Suppose a study has 80% probability of reaching statistical significance. In that case, the SNR is 2.8 and the sampling distribution of the z-value is normal with mean 2.8 and unit variance. Now, both $z = 0.8$ and $z = 4.8$ can easily occur because both are two standard deviations from the mean. These z-values correspond to $p = 0.42$ and $p = 1.6 \cdot 10^{-6}$, respectively. Many would consider these p-values to be worlds apart.

3.2. A model for the distribution of signal-to-noise ratios in a corpus

If a study has a large signal-to-noise ratio (SNR) it means that the statistical uncertainty of the effect estimate is relatively small compared to the size of the effect. Conversely, a small SNR means that the study is "noisy" in the sense that the statistical uncertainty makes it difficult to discern

the effect. To study the quality of the studies in a particular corpus, we will estimate the (latent) distribution of the SNRs.

To do this, we will extend the model of Stephens (2017), van Zwet, Schwab, and Senn (2021), and van Zwet and Gelman (2022) to account for the truncation and publication bias which are conspicuously present in some of our datasets.

Our assumptions imply that z-values are equal to SNRs plus independent standard normal errors. If we would observe the z-values without selection, we can estimate their distribution. We can then derive the distribution of the SNRs by a process of *deconvolution*, that is, by removing the standard normal error component (Efron 2016; Stephens 2017).

We construct the likelihood in four stages. First, we introduce a mixture distribution for the absolute z-values. Then, we add a parameter to account for publication bias. Next, we account for the fact that reported values are sometimes truncated. Finally, we weight the observations according to the number of z-values per study.

Our main assumption is that the distribution of the absolute z-values is well represented by a mixture of half-normal distributions. We will informally verify this assumption by inspection of the observed histograms. Noting that this assumption is equivalent to modeling the signed z-values as a mixture of zero-mean normal distributions, we have the mixture density,

$$f(z \mid p, \sigma) = \sum_{i=1}^{k} p_i \, \frac{1}{\sigma_i} \, \phi\left(\frac{z}{\sigma_i}\right), \tag{2}$$

where $\phi(\cdot)$ is the standard normal density, $p = (p_1, \ldots, p_k)$ are the non-negative mixture weights that sum to unity and $\sigma = (\sigma_1, \ldots, \sigma_k)$ are the standard deviations. These standard deviations are at least 1 because we know that z-values have a standard normal noise component. We will use $k = 4$, as we have verified that larger values of k yield nearly identical results.

The deconvolution to obtain the distribution of the SNRs is easy. We simply subtract 1 from the variances of the mixture components. So,

$$f_{\text{SNR}}(x \mid p, \sigma) = f\left(x \mid p, \sqrt{\sigma^2 - 1}\right). \tag{3}$$

To account for publication bias, we introduce a simple selection component in the likelihood. We follow Hedges (1984, 1992) and define ω as the relative risk that a result with $|z| < 1.96$ is observed compared to a result with $|z| \geq 1.96$:

$$\omega \;=\; \frac{\Pr(\text{result is published} \mid |z| < 1.96)}{\Pr(\text{result is published} \mid |z| \geq 1.96)}. \tag{4}$$

The probability of publication is

$$C(p, \sigma, \omega) = \omega \Pr(|z| < 1.96 \mid p, \sigma) + \Pr(|z| \geq 1.96 \mid p, \sigma) \tag{5}$$

and the density of the z-values conditional on publication becomes

$$f(z \mid p, \sigma, \omega, \text{pub}) = \begin{cases} \omega \, f(z \mid p, \sigma)/C(p, \sigma, \omega) & |z| \leq 1.96 \\ f(z \mid p, \sigma)/C(p, \sigma, \omega) & \text{otherwise.} \end{cases} \tag{6}$$

The z-value is sometimes left or right truncated, most often at 1.96, but also at, for example, 2.57 (when papers report $p < 0.01$), 3.29 ($p < 0.001$), or 1.64 ($p < 0.1$). To take this into account, we introduce a censoring indicator δ taking the values $<$, $>$, or $=$ for left censoring, right censoring,

or no censoring, respectively. We denote the (possibly) censored value of the z-value by $\tilde{z}$. The distribution of the pair $(\delta, \tilde{z})$ conditional on publication is

$$f(\tilde{z}, \delta \mid p, \sigma, \omega, \text{pub}) =$$
$$\begin{cases} \omega\, f(\tilde{z} \mid p, \sigma)/C(p, \sigma, \omega) & \delta = \texttt{"="},\ \tilde{z} < 1.96, \\ f(\tilde{z} \mid p, \sigma)/C(p, \sigma, \omega) & \delta = \texttt{"="},\ \tilde{z} \geq 1.96, \\ \omega\, F(\tilde{z} \mid p, \sigma)/C(p, \sigma, \omega) & \delta = \texttt{"<"},\ \tilde{z} < 1.96, \\ (\omega\, F(1.96 \mid p, \sigma) + F(\tilde{z} \mid p, \sigma) - F(1.96 \mid p, \sigma))/C(p, \sigma, \omega) & \delta = \texttt{"<"},\ \tilde{z} \geq 1.96, \\ (1 - F(\tilde{z} \mid p, \sigma))/C(p, \sigma, \omega) & \delta = \texttt{">"},\ \tilde{z} \geq 1.96, \\ ((1 - F(1.96 \mid p, \sigma)) + \omega\,(F(1.96 \mid p, \sigma) - F(\tilde{z} \mid p, \sigma)))/C(p, \sigma, \omega) & \delta = \texttt{">"},\ \tilde{z} < 1.96. \end{cases}$$

The first two rows cover the case without truncation. Rows three and four are left censoring (statements such as $p > 0.05$), and rows five and six deal with right censoring (e.g., $p < 0.001$). We also introduce artificial censoring when z-values are reported to be exactly zero, by treating them as $|z| < 0.5$.

Most of the datasets have multiple z-values per study. To prevent studies with many z-values from dominating the likelihood, we weight observations by w_j which is the inverse of the number of z-values originating from study j. This does not account for the dependence among z-values from the same study but that does not concern us because we do not attempt to quantify the uncertainty of our estimates. The reason is that our datasets are so large that the sampling uncertainty which determines standard errors and confidence intervals, is negligible compared to model uncertainty. The full weighted log likelihood is

$$\ell(p, \sigma, \omega) = \sum_{j=1}^{n} w_j \log f(\tilde{z}_j, \delta_j \mid p, \sigma, \omega, \text{pub}). \tag{7}$$

Even for large datasets, numerical optimization of this log likelihood is feasible, taking just minutes. For very large collections of z-values, we cap the sample size at 50,000, using study-level thinning as described in the Data section.

The estimated distribution of the z-values without publication bias is $f(z \mid \hat{p}, \hat{\sigma})$ as defined in (2) and consequently the estimated distribution of the SNRs is

$$f_{\text{SNR}}(x \mid \hat{p}, \hat{\sigma}) = f\left(x \mid \hat{p}, \sqrt{\hat{\sigma}^2 - 1}\right). \tag{8}$$

We can derive many important quantities from the distribution of the SNRs. For example, the PoS is a function of the SNR, so we can obtain its distribution by transforming the distribution of the SNRs. In particular, the proportion of studies that have at least 80% probability of reaching statistical significance is

$$\Pr(\text{PoS} \geq 0.8) = \Pr(|\text{SNR}| \geq 2.8). \tag{9}$$

The marginal density of the SNRs, together with the conditional density of the z-values given the SNRs, determines the joint distribution. And from the joint distribution we can derive the conditional distribution of the SNR given the z-value. We provide the formulas in Appendix A. In turn, this allows us to derive the conditional probability of the correct sign,

$$\text{Probability of the correct sign} = \Pr(z \cdot \text{SNR} > 0 \,\big|\, |z|), \tag{10}$$

Finally, we imagine that a second replication study with exactly the same SNR yields another z-value z_{repl}. This z_{repl} is conditionally independent of the original z given SNR, and has the same distribution. Thus van Zwet and Goodman (2022) were able to derive the conditional distribution of $|z_{\text{repl}}|$ given $|z|$. In particular, we obtain

$$\text{Probability of "successful replication"} = \Pr(z_{\text{repl}} \cdot z > 0 \text{ and } |z_{\text{repl}}| > 1.96 \,\big|\, |z|) \tag{11}$$

All the computations that yielded the results reported in the next section are documented in the online supplement to this paper.

4. Results

4.1. Distributions of z-values

Figure 1 displays the histograms of the observed absolute z-statistics of our 8 curated sets and 6 representative domain samples, arranged by their estimated assurances $\overline{\text{PoS}}$. The solid curves are the fitted mixtures of half-normal distributions conditional on publication. The discontinuities ("jumps") of the densities at $|z| = 1.96$ are a feature of Hedges' selection model. The dashed curves are the "corrected" unconditional distributions. We find that the estimated distributions generally track the histograms well and conclude that our model is appropriate for our purpose of quantifying large-scale properties of research areas.

Figure 3 compares the distributions of the observed z-values of the original studies and their replications from SCORE and OSC replication projects (Open Science Collaboration 2015; Tyner et al. 2026).

Mixtures of half-normal densities are necessarily decreasing. Apart from the effects of publication bias, we also observe this feature in the histograms. We claim that this is typical of large collections of z-values. To further support this claim, we have broken down the Cochrane data by 19 medical specialties and the data from `clinicaltrials.gov` by trial phase; see Figure S2. Again, we see decreasing distributions of absolute z-values.

Another common feature of the histograms is the selection on statistical significance which is especially severe in the 6 domain samples. Apart from the selection on significance, the main difference between the distributions of the z-values is their width. A wider distribution of z-values implies a wider distribution of SNRs. In other words, the studies in that research area tend to have larger absolute SNRs; they tend to be less noisy. Since the PoS is an increasing function of the absolute SNR, this also means that a higher proportion of results will reach statistical significance.

4.2. Quantitative results

We report the quantitative characteristics of the corpora in Table 2. The first column has the estimated Hedges' relative risk of publication $\hat{\omega}$, as defined in (4). We find a clear divide between studies in terms of these relative risks, which we can also observe in Figure 1. We find much less selection on significance in the curated datasets, with $\hat{\omega} \geq 0.64$ across all eight sets, than in the domain-representative samples, where $\hat{\omega}$ ranges from 0.08 for Barnett and Wren (2019) to 0.56 for clinical trial registries. However, we must be careful not to over-interpret this result. The corpus of Barnett and Wren (2019) is based on confidence intervals, and it is possible that non-significant results were actually reported, just not with the associated confidence intervals.

In light of the replication crisis in science, it is tempting to focus on these estimates of the extent of publication bias. However, there are several reasons for caution. Many studies in our datasets have multiple outcomes, and we do not know which of them drive publication decisions. Moreover, while

Hedges' single-parameter selection model is useful for adjusting our estimates of the distributions of the z-statistics, it seems too limited for reliably quantifying and comparing the extent of publication bias across fields of research. This is also not the goal of our study.

The next two columns in Table 2 which are labeled "Significance" report proportions of significant results in the 16 corpora before and after adjustment for publication bias. We may refer to the adjusted proportion as the "assurance" and denote it by $\overline{\text{PoS}}$, as defined in (1). The adjusted proportion is strictly less than the raw proportion, unless $\hat{\omega} = 1$ which would mean that there is no discernible publication bias. Across the 16 corpora, the estimated assurances range from 0.28 to 0.52. The estimated assurance of phase 3 clinical trials in Figure S2 matches the estimated success rates (probability of moving to regulatory approval following trial) in phase 3 trials in drug development reported by Thomas et al. (2021).

By transforming the distribution of the signal-to-noise ratios, we obtain the full distribution of the PoS within each dataset. The left panel of Figure 2 displays the complementary cumulative distribution functions (survival functions) of the PoS in the 16 datasets. In other words, it shows what proportions of studies in each dataset "survive" to achieve a given level of power. We find that between 12% and 33% of studies achieve 80% PoS; see Table S2.

Since the great majority of studies are targeted at reaching 80% power, the proportion of studies that reach 80% PoS may be tentatively interpreted as the proportions of studies where the assumptions of the sample size calculation were true. In general, this proportion is low. However, sample size calculations are not supposed to be aimed at the true effect (which is never known) but at the effect one would not want to miss (Senn 2002). In that sense, a low proportion of studies with 80% PoS may simply mean that it is hard to come up with interventions that have meaningful effects. Pressures of time, money, and logistics also drive down sample sizes.

The three columns in Table 2 that are labeled "Successful replication" report what would happen if a randomly selected study from a particular corpus would be replicated exactly. Exact replications are not possible in practice, so we should think of these as "in silico" replications. We show the probability of obtaining a statistically significant result in the same direction both unconditionally and conditionally on the original study having $|z| = 1.96$ or $|z| \geq 1.96$; see equation (11). We find that if a result is marginally statistically significant at the 5% level (i.e., $|z| = 1.96$) then the probability of "successful replication" is never much more than the base rate, and often even less. This will be a surprise to those who think of statistical significance as a stand-in for replication. This phenomenon is *not* a consequence of many studies having low SNRs (low PoS). If all the studies had higher SNRs, then the assurance would be higher but observing $|z| = 1.96$ would be even less impressive.

Finally, the three columns that are labeled "Correct sign" report the estimated proportion of studies where the observed effect has the correct direction (sign). Again, we also condition on $|z| = 1.96$ or $|z| \geq 1.96$, as defined in equation (10). If we focus on the case where $|z| = 1.96$ then we find a large difference between the probabilities of successful replication and the correct sign. Across the 16 corpora, the estimated probability of replication ranges from 30% to 44% while the estimated probability of correct sign ranges from 85% to 97%.

The right panel of Figure 2 displays the probabilities of successful replication and correct sign conditional on the observed absolute z-value; see also Figure S1. The probability of correct sign is always much higher than the probability of successful replication.

Figure 3 shows the z-values of the studies that were selected for replication in the Open Science Collaboration project in psychology (Open Science Collaboration 2015) and the SCORE project in the social and behavioral sciences (Tyner et al. 2026). We also show the z-values of the replications. The studies were selected to be representative of the respective fields, and we note the strong selection on significance. Without that selection, the z-values of the replications show a decreasing density.

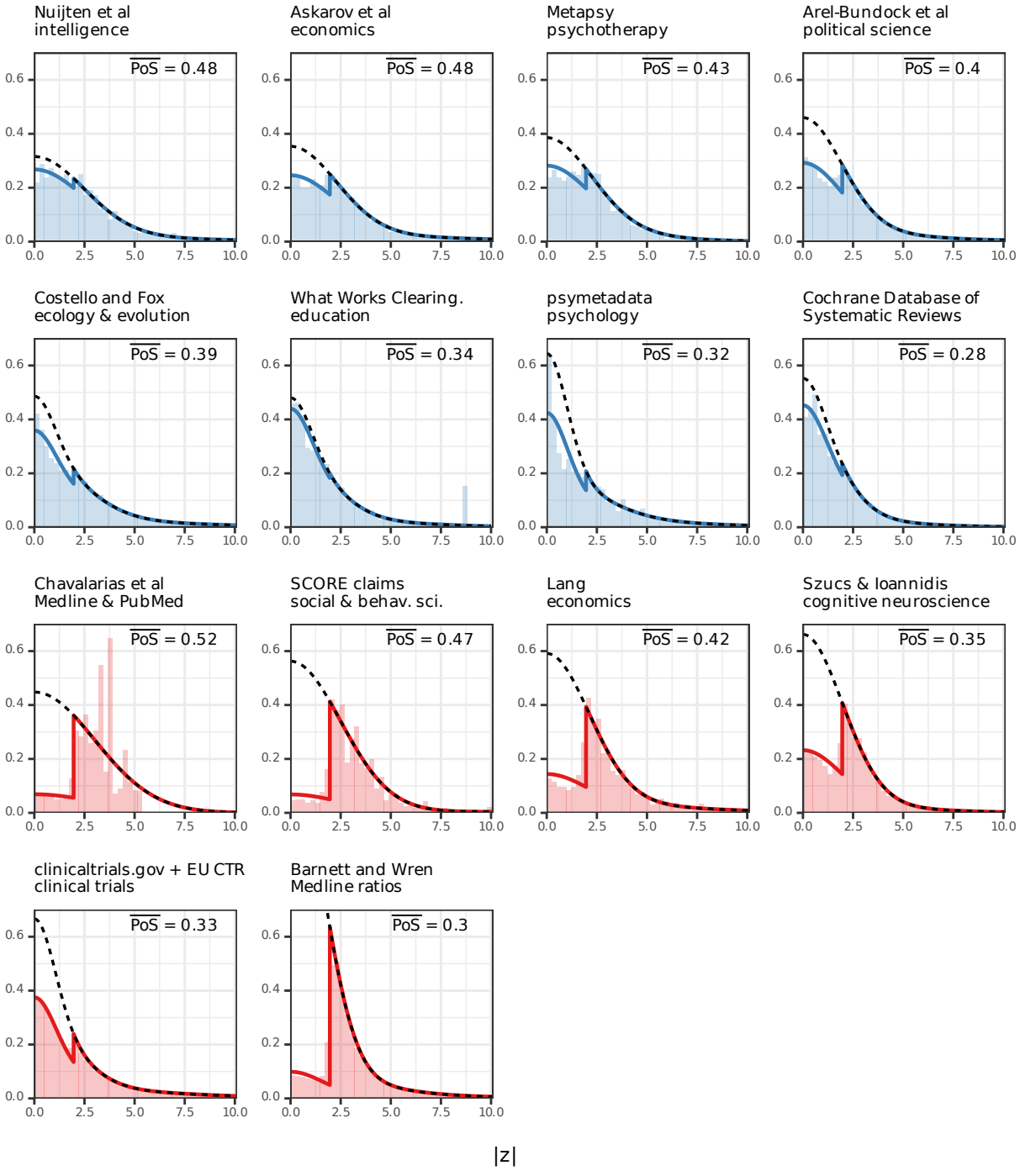

Figure 1: Histograms of absolute z-values across 14 non-replication corpora with fitted curves overlaid on top. The solid curves are densities, fitted using mixtures of half-normal distributions and conditioning on publication. Dashed curves are the unconditional "corrected" distributions, scaled to match the solid curve in region $|z| > 1.96$. Blue: curated sets; Red: representative domain samples.

		Significance		"Successful" replication			Correct sign		
					Conditional on			Conditional on	
Corpus	$\hat{\omega}$	signif.	$\overline{PoS}$	uncond.	$\lvert z\rvert = 1.96$	$\lvert z\rvert \geq 1.96$	uncond.	$\lvert z\rvert = 1.96$	$\lvert z\rvert \geq 1.96$
Curated sets									
Nuijten et al.	0.85	0.53	0.48	0.47	0.41	0.75	0.88	0.96	0.99
Askarov et al.	0.70	0.52	0.48	0.46	0.40	0.75	0.88	0.96	0.99
Metapsy	0.73	0.48	0.43	0.41	0.40	0.70	0.87	0.96	0.99
Arel-Bundock et al.	0.64	0.47	0.40	0.38	0.36	0.68	0.85	0.96	0.99
Costello and Fox	0.74	0.42	0.39	0.37	0.35	0.74	0.79	0.89	0.98
What Works Clearinghouse	0.92	0.34	0.34	0.32	0.32	0.70	0.77	0.88	0.97
psymetadata	0.66	0.36	0.32	0.30	0.30	0.74	0.72	0.81	0.96
Cochrane	0.82	0.33	0.28	0.26	0.30	0.60	0.77	0.90	0.97
Representative domain samples									
Chavalarias et al.	0.15	0.87	0.52	0.50	0.44	0.76	0.89	0.97	1.00
SCORE, all claims	0.12	0.87	0.47	0.45	0.41	0.73	0.88	0.96	0.99
Lang	0.24	0.58	0.42	0.41	0.38	0.70	0.86	0.96	0.99
Szucs and Ioannidis	0.35	0.62	0.35	0.33	0.36	0.62	0.84	0.96	0.99
clinicaltrials.gov + EU CTR	0.56	0.44	0.33	0.32	0.30	0.73	0.74	0.85	0.97
Barnett and Wren	0.08	0.84	0.30	0.28	0.30	0.58	0.81	0.94	0.98
Replication efforts									
SCORE replications	0.70	0.64	0.51	0.49	0.39	0.78	0.88	0.96	0.99
OpenSciCollab replications	1.05	0.35	0.36	0.35	0.31	0.74	0.80	0.91	0.98

Table 2: Summary of signal-to-noise ratio modeling results for the 16 corpora used in the paper. Rows are grouped into curated sets, representative domain samples, and replication projects. Calculations of power, replication, and direction of effects use the distributions of z-values and signal-to-noise ratios by fitted mixture models without selection.

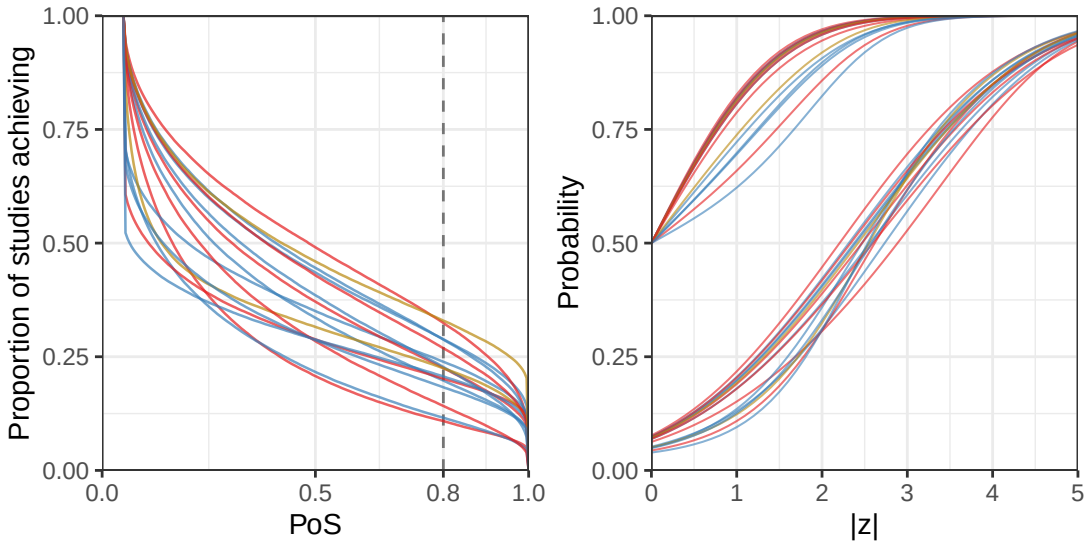

Figure 2: Left panel: "survival curves" of the probability of significance: 1 − (CDF of PoS) for each of the 16 corpora. Right panel: conditional probabilities that the observed direction of the effect matches the true direction (top set of curves) and the conditional probability of "successful" replication (bottom set). Colors distinguish curated sets, representative domain samples, and replication efforts.

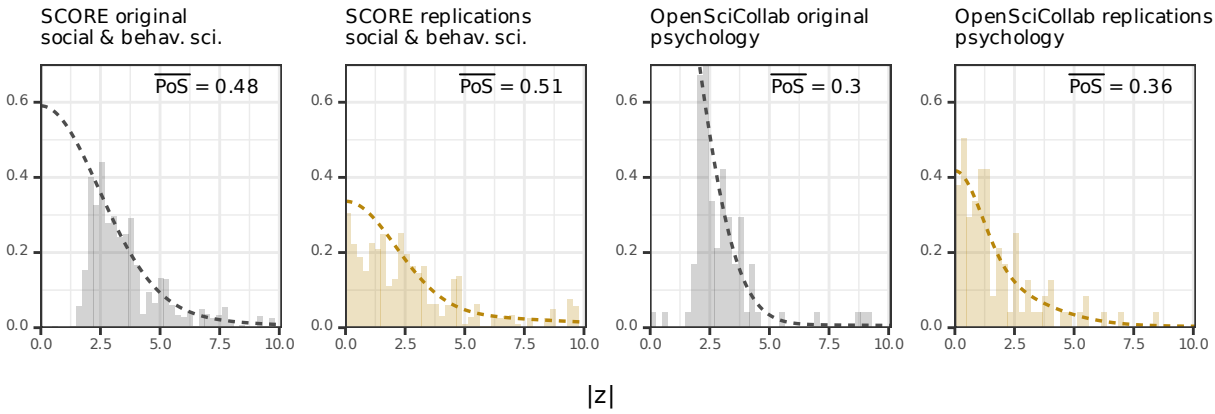

Figure 3: Histograms of absolute z-values for the SCORE and OSC original-study and replication-study distributions. Only the unconditional "corrected" mixture density is shown. Gray panels show original studies; yellow panels show replication studies.

5. Discussion

5.1. An alternative to the binary model of scientific discovery

The 2×2 model (Ioannidis 2005) is often taken to be a reasonable, if simplified, description of scientific research. According to this model, the probabilities of significance (PoS) of the "true relationships" and "no relationships" are well separated. The PoS of "no relationships" is 5% by the design of the statistical test, and the PoS of "relationships" is (or, at least, should be) much larger by the design of the experiment. If the ratio of "true relationships" to "no relationships" is not too low, then we can reliably separate them. That is, the positive predictive value (the conditional probability of a true relationship given statistical significance) and the negative predictive value (the conditional probability of no relationship given non-significance) are both high. Moreover, the conditional probability of successful replication (reaching statistical significance again in an exact replication study) is also high. Failure to replicate may then be taken as strong evidence that there is actually no relationship.

We have developed an alternative statistical model that goes beyond the 2×2 framework by allowing the PoS to range continuously from 0.05 to 1. We allow for truncation and used Hedges' model to adjust for selective reporting. We fitted our model to sixteen corpora across diverse fields collected from publicly available sources; see Figures 1 and S2.

5.2. What our meta-analysis says about effect sizes and replication rates

We discuss four features that hold across all the datasets of empirical research we collected.

First, apart from the effects of the selection on significance, the absolute z-values in all the corpora have a decreasing density. This suggests that the absolute SNRs also have a decreasing density. In other words, studies with low SNRs are more common than those with high SNRs. Why do these different corpora, coming from many different experiments in different fields, show similar distributions of the absolute z-values? We tentatively ascribe this to a sort of anthropic principle (Gelman 2018) whereby researchers respond to the pressures of time, money, and logistics on one side and the goal to reach statistical significance on the other, by making their studies just about large enough to have a fair chance at reaching statistical significance, but not larger. We do find

that there are considerable differences between research domains in the widths of these distributions which means that the studies in some areas tend to have higher SNRs (be less noisy) than in other areas.

Second, the probability of significance (PoS) tends to be low and only a small minority of studies achieve 80% power. The effect of a treatment or intervention relative to the standard deviation of the outcome is often small. To reach a PoS of 80% would require a sample size which is larger than what is available or affordable. Under such circumstances, researchers will use some optimistic effect size for their sample size calculation, and go ahead with the study. We agree that it may be better to do a small study than to do no study at all.

Third, reaching statistical significance does not imply a high probability of replication; see Figure 2 and Figure S1. Even at $p = 0.001$ ($|z| = 3.29$), statistically significant replication is far from assured. This is not difficult to understand. The majority of studies have low PoS, but, for example, 1 in 5 studies with a PoS of only 20% will reach statistical significance. Conversely, many studies that have reached statistical significance have a low PoS and will not reach statistical significance upon replication.

Fourth, in the absence of selection, statistical significance *does* implies that the sign of the effect is likely correct. In other words, the error of rejecting the null hypothesis in the wrong direction is unlikely. Directionally wrong conclusions from statistically significant results are relatively rare. If that were be the case, scientific progress would at best be slow.

These conclusions are especially relevant for a hypothetical new study that is comparable to those in our corpora. If such a study is preregistered and no questionable research practices are committed, then a z-value in the range of 2 to 3, would imply roughly a 90% chance of getting the correct sign. Then, if that new study were repeated, we would expect about 50% probability of reaching statistical significance again; see Figure 2.

Unfortunately, much research is not preregistered and publication bias, researcher degrees of freedom, paper mills and outright fraud are major concerns. So, we are *not* saying that any statistically significant result that appears in print has a high probability of pointing in the right direction. If we have good reasons to doubt the study design or integrity of the analysis, we should use that information. In other words, not every result should be treated as a typical draw from our corpora. Our qualified optimism extends only to research for which we do not have decisive reasons to distrust.

5.3. Reasons for qualified optimism about empirical research

The standard practice in empirical research is that if a result reaches a statistical significance threshold, it is treated as real and the corresponding estimate is taken at face value. Results that are not significant are presented as a null findings or, at best, scenarios where there is not enough information to draw any conclusions. It is then considered scandalous when a statistically significant published result does not replicate (Tversky and Kahneman 1971).

One way to summarize our results is to say that empirical research often leads to true findings—in the sense of correctly identifying the direction of an effect—that do not replicate in the sense of achieving statistically significant result in the same direction as the original study. Therefore, failure to replicate should not be taken to mean that the original result was a fluke or a fake. The pessimism in much of the meta-scientific literature reflects a one-sided interpretation of Figure 2 and Figure S1 focusing only on the low replication rates. This skewed view has serious consequences as it may lead to dangerous nihilism among policy makers (Garisto 2025). With our study, we hope to rectify it.

We emphasize one last time that our optimism does not extend to studies where we have reason for distrust.

6. Acknowledgments

We acknowledge the creators of the corpora and the authors of the original studies for their great effort. In particular, we are grateful to Frantisek Bartos, Ryan Briggs, Chris Doucouliagos, Chris Mead, Martina Sladekova, and Tom Stanley for helping us understand the datasets. The conclusions we draw from the datasets are entirely our own. We thank the U.S. Office of Naval Research for partial support of this work.

References

Aquarius, René, Elisabeth M. Bik, David Bimler, Morten P. Oksvold, and Kevin Patrick (2025). "Tackling paper mills requires us to prevent future contamination and clean up the past—the case of the journal Bioengineered." *Bioengineered* 16.1, p. 2542668.

Arel-Bundock, Vincent, Ryan C. Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and Tom D. Stanley (2022). "Quantitative political science research is greatly underpowered." *I4R Discussion Paper Series* 6.

Askarov, Zohid, Anthony Doucouliagos, Hristos Doucouliagos, and T. D. Stanley (2023). "The significance of data-sharing policy." *Journal of the European Economic Association* 21.3, pp. 1191–1226.

Bak-Coleman, Joseph, Richard Mann, Carl Bergstrom, Kevin Gross, and Jevin West (2022). "Revisiting the replication crisis without false positives." *Center for Open Science*. `https://osf.io/preprints/socarxiv/rkyf7_v1/`.

Barnett, Adrian Gerard and Jonathan D. Wren (2019). "Examination of CIs in health and medical journals from 1976 to 2019: an observational study." *BMJ Open* 9.11, e032506.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. (2016). "Evaluating replicability of laboratory experiments in economics." *Science* 351.6280, pp. 1433–1436.

Chavalarias, David, Joshua David Wallach, Alvin Ho Ting Li, and John P. A. Ioannidis (2016). "Evolution of reporting P values in the biomedical literature, 1990-2015." *Journal of the American Medical Association* 315.11, pp. 1141–1148.

Cochrane Collaboration (2025). *Cochrane Database of Systematic Reviews*. URL: `https://www.cochranelibrary.com/cdsr/about-cdsr/`.

Costello, Laura and Jeremy W. Fox (2022). "Decline effects are rare in ecology." *Ecology* 103.6, e3680.

Cuijpers, Pim, Mathias Harrer, Clara Miguel, et al. (2025). "Cognitive behavior therapy for mental disorders in adults: A unified series of meta-analyses." *JAMA Psychiatry* 82.6, pp. 563–571.

Efron, Bradley (2016). "Empirical Bayes deconvolution estimates." *Biometrika* 103.1, pp. 1–20.

European Medicines Agency (2025). *EU Clinical Trials Register*. URL: `https://www.clinicaltrialsregister.eu/`.

Garisto, Dan (2025). "Trump order gives political appointees vast powers over research grants." *Nature* 644.8077, pp. 585–586.

Gelman, Andrew (2018). "The anthropic principle in statistics." *Statistical Modeling, Causal Inference, and Social Science*. 23 May. `https://statmodeling.stat.columbia.edu/2018/05/23/anthropic-principle-statistics/`.

Gelman, Andrew and John Carlin (2014). "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors." *Perspectives on Psychological Science* 9.6, pp. 641–651.

Gelman, Andrew and Francis Tuerlinckx (2000). "Type S error rates for classical and Bayesian single and multiple comparison procedures." *Computational Statistics* 15.3, pp. 373–390.

Greenland, Sander (2017). "The need for cognitive science in methodology." *American Journal of Epidemiology* 186.6, pp. 639–645.

Harrer, Mathias, Eirini Karyotaki, and Pim Cuijpers (2022). *metapsyData: Meta-Analytic Psychotherapy Databases*. R package. URL: https://CRAN.R-project.org/package=metapsyData/.

Hedges, Larry V. (1984). "Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences." *Journal of Educational Statistics* 9.1, pp. 61–85.

Hedges, Larry V. (1992). "Modeling publication selection effects in meta-analysis." *Statistical Science* 7.2, pp. 246–255.

Herold, R. (2025). "Aggregating and analysing clinical trials data from multiple public registers using R package ctrdata." *Research Synthesis Methods*, pp. 1–33. DOI: 10.1017/rsm.2025.10061.

Institute of Education Sciences (2025). *What Works Clearinghouse Study Findings*. URL: https://ies.ed.gov/ncee/wwc/studyfindings/.

Ioannidis, John P. A. (2005). "Why most published research findings are false." *PLoS Medicine* 2.8, e124.

Lang, Kevin (2025). "How credible is the credibility revolution?" *Journal of Labor Economics* 43, pp. 635–663.

Matusz, Pawel J., Anna Abalkina, and Dorothy V. M. Bishop (2025). "The threat of paper mills to biomedical and social science journals: The case of the Tanu.pro paper mill in *Mind, Brain, and Education*." *Mind, Brain, and Education* 19.2, pp. 90–100.

National Library of Medicine (2025). *ClinicalTrials.gov*. URL: https://clinicaltrials.gov/.

Neves, Kleber, Pedro B. Tan, and Olavo B. Amaral (2022). "Are most published research findings false in a continuous universe?" *Plos One* 17.12, e0277935.

Nuijten, Michèle B., Marcel A. L. M. Van Assen, Hilde E. M. Augusteijn, Elise A. V. Crompvoets, and Jelte M. Wicherts (2020). "Effect sizes, power, and biases in intelligence research: A meta-meta-analysis." *Journal of Intelligence* 8.4, p. 36.

Open Science Collaboration (2015). "Estimating the reproducibility of psychological science." *Science* 349.6251, aac4716.

Rodriguez, Josue E. and Donald R. Williams (2022). "psymetadata: An R package containing open datasets from meta-analyses in psychology." *Journal of Open Psychology Data*. DOI: 10.5334/jopd.61/.

Schwab, Simon (2024). *cochrane: Import data from the Cochrane database of systematic reviews (CDSR)*. R package. URL: https://github.com/schw4b/cochrane/.

Senn, Stephen (2002). *Cross-over Trials in Clinical Research*. Wiley.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011). "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological Science* 22.11, pp. 1359–1366.

Stephens, Matthew (2017). "False discovery rates: a new deal." *Biostatistics* 18.2, pp. 275–294.

Szucs, Denes and John P. A. Ioannidis (2017). "Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature." *PLoS Biology* 15, e2000797.

Thomas, David, Daniel Chancellor, Amanda Micklus, Sara LaFever, Michael Hay, Shomesh Chaudhuri, Robert Bowden, and Andrew W. Lo (2021). *Clinical Development Success Rates and Contributing Factors, 2011–2020*. Report. Biotechnology Innovation Organization (BIO). URL: https://go.bio.org/rs/490-EHZ-999/images/ClinicalDevelopmentSuccessRates2011_2020.pdf.

Tversky, Amos and Daniel Kahneman (1971). "Belief in the law of small numbers." *Psychological Bulletin* 76, pp. 105–110.

Tyner, Andrew H. et al. (2026). "Investigating the replicability of the social and behavioural sciences." *Nature* 652, pp. 143–150.

van Zwet, Erik and Andrew Gelman (2022). "A proposal for informative default priors scaled by the standard error of estimates." *American Statistician* 76.1, pp. 1–9.

van Zwet, Erik and Steven N. Goodman (2022). "How large should the next study be? Predictive power and sample size requirements for replication studies." *Statistics in Medicine* 41.16, pp. 3090–3101.

van Zwet, Erik, Simon Schwab, and Stephen Senn (2021). "The statistical properties of RCTs and a proposal for shrinkage." *Statistics in Medicine* 40.27, pp. 6107–6117.

Yang, Yefeng, Alfredo Sánchez-Tójar, Rose E. O'Dea, Daniel W. A. Noble, Julia Koricheva, Michael D. Jennions, Timothy H. Parker, Malgorzata Lagisz, and Shinichi Nakagawa (2023). "Publication bias impacts on effect size, statistical power, and magnitude (type M) and sign (type S) errors in ecology and evolutionary biology." *BMC Biology* 21.1, p. 71.

Yang, Yefeng, Erik van Zwet, Nikolaos Ignatiadis, and Shinichi Nakagawa (2024). "A large-scale in silico replication of ecological and evolutionary studies." *Nature Ecology & Evolution*, pp. 1–5.

A. Details of calculations

Suppose the marginal distribution of the z-value is a mixture of zero-mean normal distributions,

$$f(z \mid p, \sigma) = \sum_{i=1}^{k} p_i \, \frac{1}{\sigma_i} \, \phi\!\left(\frac{z}{\sigma_i}\right).$$

Suppose also that the conditional distribution of the z-value, given the SNR, is normal with mean SNR and variance 1. Together, the marginal and conditional distributions determine the joint distribution of the z-value and the SNR. Therefore, we also have the conditional distribution of the SNR given the z-value. It is a mixture of normal distributions with weights $q_i(z)$, means $\mu_i(z)$, standard deviations τ_i given by

$$q_i(z) = \frac{p_i \phi(z/\sigma_i)/\sigma_i}{f(z \mid p, \sigma)}, \quad \mu_i(z) = \frac{\sigma_i^2 - 1}{\sigma_i^2} \, z \quad \text{and} \quad \tau_i = \frac{\sqrt{\sigma_i^2 - 1}}{\sigma_i}. \tag{12}$$

The conditional probability of the correct sign is

$$\Pr(z \cdot \mathrm{SNR} > 0 \mid |z|) = \sum_{i=1}^{k} q_i(z) \Phi\left(\frac{\mu_i(|z|)}{\tau_i}\right). \tag{13}$$

If z_{repl} is the z-value of an exact replication study, then the conditional distribution of z_{repl} given the z-value of the original study is a mixture of normal distributions with weights $q_i(z)$, means $\mu_i(z)$ and variances $\tau_i^2 + 1$. The conditional probability of "successful replication" is

$$\Pr(z_{\mathrm{repl}} \cdot z > 0 \text{ and } |z_{\mathrm{repl}}| > 1.96 \mid |z|) = \sum_{i=1}^{k} q_i(z) \Phi\left(\frac{\mu_i(|z|) - 1.96}{\sqrt{\tau_i^2 + 1}}\right). \tag{14}$$

		Significance		PoS at least 80%		
					Conditional on	
Corpus	$\hat{\omega}$	signif.	$\overline{PoS}$	uncond.	$\lvert z\rvert = 1.96$	$\lvert z\rvert \geq 1.96$
Curated sets						
Nuijten et al.	0.85	0.53	0.48	0.29	0.11	0.57
Askarov et al.	0.70	0.52	0.48	0.29	0.10	0.58
Metapsy	0.73	0.48	0.43	0.22	0.09	0.49
Arel-Bundock et al.	0.64	0.47	0.40	0.20	0.07	0.48
Costello and Fox	0.74	0.42	0.39	0.24	0.09	0.59
What Works Clearinghouse	0.92	0.34	0.34	0.18	0.07	0.52
psymetadata	0.66	0.36	0.32	0.20	0.08	0.61
Cochrane	0.82	0.33	0.28	0.12	0.05	0.38
Representative domain samples						
Chavalarias et al.	0.15	0.87	0.52	0.32	0.13	0.60
SCORE, all claims	0.12	0.87	0.47	0.27	0.10	0.54
Lang	0.24	0.58	0.42	0.23	0.08	0.51
Szucs and Ioannidis	0.35	0.62	0.35	0.14	0.06	0.38
clinicaltrials.gov + EU CTR	0.56	0.44	0.33	0.20	0.07	0.59
Barnett and Wren	0.08	0.84	0.30	0.11	0.04	0.35
Replication efforts						
SCORE replications	0.70	0.64	0.51	0.33	0.09	0.64
OpenSciCollab replications	1.05	0.35	0.36	0.23	0.08	0.60

Table S2: Summary of probability-of-significance results for the 16 corpora used in the paper. Rows are grouped into curated sets, representative domain samples, and replication projects.

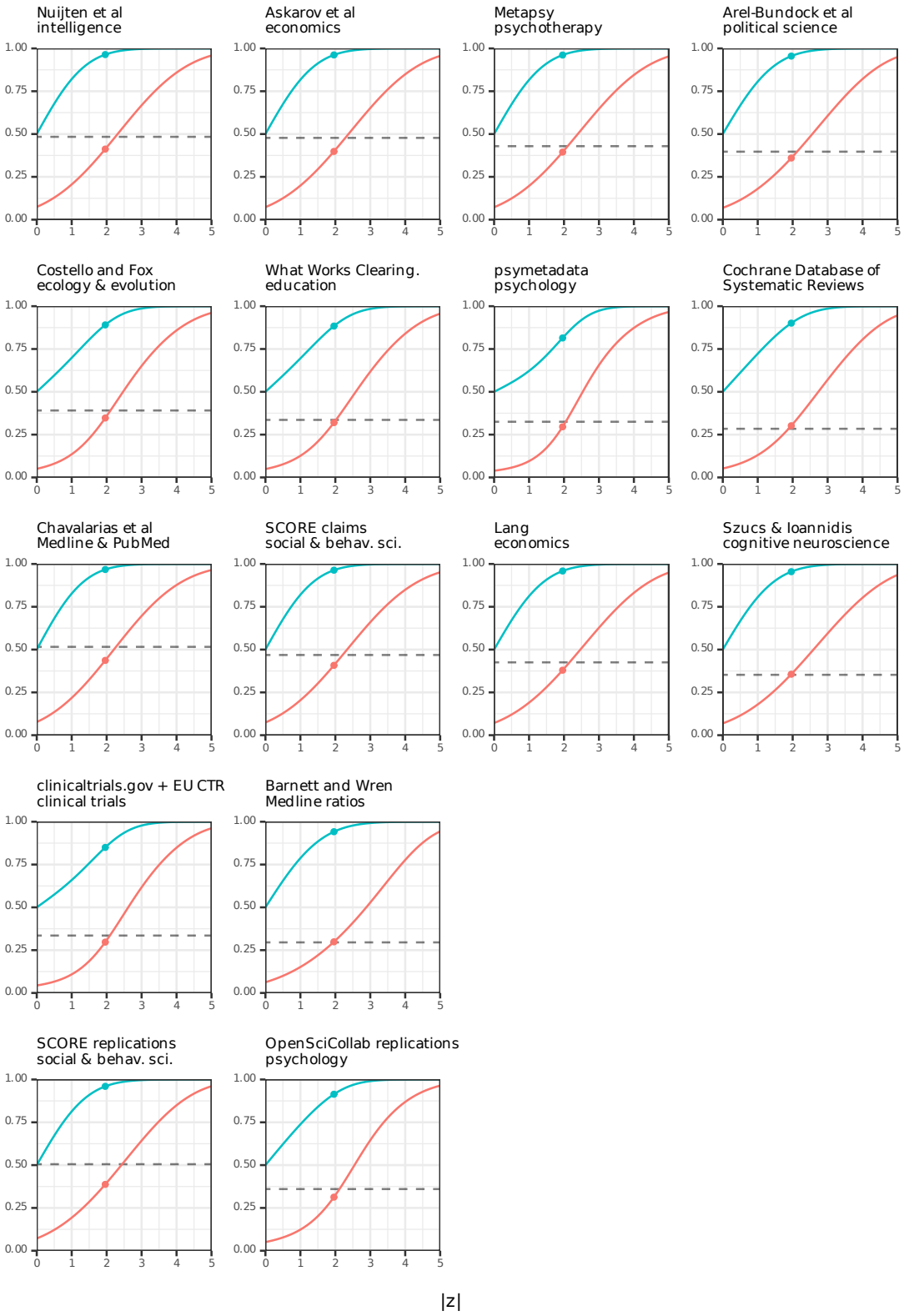

Figure S1: Relationship between absolute z-values and the probability that the direction of the observed effect matches the true direction (blue curves) and the probability of "successful replication" (red curves). The vertical line is $\overline{\text{PoS}}$.

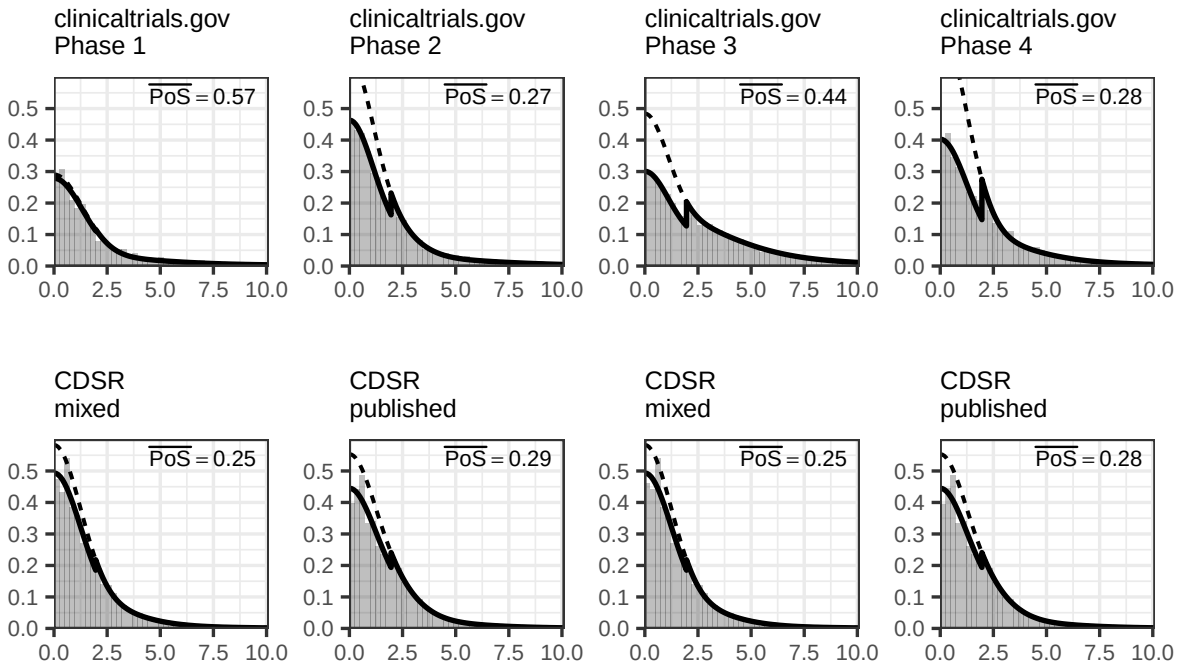

Figure S2: Subsets of interventional treatment studies from `clinicaltrials.gov`, broken down by phase (top row) and from Cochrane database (bottom row), grouped by specialty (out of 19 included in the dataset): two with lowest $\overline{\text{PoS}}$ on the left and two with the highest on the right. The peaks in the histograms at $|z| = 1.96, 2.58, 3.29$ are due to trials reporting (two-sided) $p < 0.05, p < 0.01, p < 0.001$.

B. Datasets included in BEAR

This appendix describes the 16 corpora used in the paper. The full BEAR database contains additional datasets. More complete documentation is available online.

B.1. Nuijten et al. 2020 and psymetadata (psychology)

Reference: A set of datasets compiled by Rodriguez and Williams (2022). We also disaggregated a large meta-meta-analysis of intelligence research (Nuijten et al. 2020).

Snapshot date: January 2026

Research question: *psymetadata* is an R package that curates multiple open datasets from published meta-analyses in psychology for teaching and demonstrations. Nuijten et al. (2020) is a large meta-meta-analysis of intelligence research, which focused on power and small study effects.

Data availability: 22 datasets are distributed with the `psymetadata` R package (CRAN). We do not use its Many Labs 2 dataset, which we obtain from the original source and include separately. After excluding Many Labs 2 and separately representing the intelligence meta-meta-analysis, the remaining psychology datasets range from 65 rows to over 1,000. The package is GPL-3; individual datasets originate from the cited source papers.

Data processing: Processing was minimal. We used regular expressions to extract year labels from titles. We dropped some invalid rows (missing, non-finite, or non-positive standard errors). Signed z-values were computed as effect sizes divided by their standard errors.

Study characteristics:

- **psymetadata:** the designs of the primary studies are not harmonized across the imported datasets.
- **psymetadata:** the effect-size measure is not harmonized across the imported datasets.
- **psymetadata:** records are grouped by original psymetadata dataset.
- **Nuijten:** the designs of the primary studies are not classified.
- **Nuijten:** BEAR retains the reported effect estimates and standard errors without a harmonized effect-size measure.

B.2. Askarov et al. 2023 (economics)

Reference: Askarov et al. (2023).

Research question: Impact of mandatory data sharing on (excessive) statistical significance in economics papers

Data availability: The file `Mandatory data-sharing 30 Aug 2022.dta` may be downloaded from `https://github.com/anthonydouc/Datasharing/blob/master/Stata/`; the reproducibility package is available at `https://dataverse.harvard.edu/file.xhtml?fileId=7884702&version=1.0`.

Data description and source: The paper includes 64,000 economic estimates surveyed by Ioannidis et al. (2017) plus results of "searching through numerous databases, journal websites, and also an email survey of 109 authors known to have produced meta-analyses of economic topics. We cast the widest net possible to capture empirical economic estimates. Thus, some of the papers included in these meta-analyses were originally published in political science and psychology journals. Nevertheless, these meta-studies deal with economics topics." Later they note: "we collect 166,924 parameter estimates from 345 distinct research areas, of which 20,121 were published in 24 of the leading general interest and field journals." The paper mentions 12,521 observations from before and 2,537 from after the introduction of data-sharing policies in 24 journals; the whole reproducibility dataset has 32 journals (the authors removed some smaller ones) and 22,172 rows.

Notes: The article by Askarov et al. starts from the Ioannidis-Stanley-Doucouliagos dataset, which is also used by Arel-Bundock et al. (another dataset in BEAR). Both datasets incorporate additional data and target different fields of research, but there is likely substantial overlap between the two sets.

Data processing: We use the same source file `Mandatory data-sharing 30 Aug 2022.dta` (11.6 MB) as in the reproducibility package. Signed z-values were computed as reported effect sizes divided by their standard errors.

Study characteristics:

- **Study ID:** we use the source study identifier.
- **Meta-analysis ID:** we use `filename`.
- **Topic:** we use the source's flag to distinguish macroeconomics or microeconomics/other.
- **Method:** The source flags experimental and mixed observational/experimental research. We label estimates with neither flag set as "not RCT, not mixed." This category may include quasi-experimental research, which the source does not separately identify.
- We do not have access to the measure of effect size (for example, ratio vs. mean difference).

B.3. Metapsy (psychotherapy)

Studies included in *Metapsy* (`https://www.metapsy.org/`), a set of living meta-analytic databases of randomized trials of psychotherapeutic interventions.

Snapshot date: January 2026

Data availability: Metapsy is maintained by an international collaboration led by Vrije Universiteit Amsterdam. Databases are available at `https://www.metapsy.org/` and through the R package `metapsyData` (Harrer, Karyotaki, and Cuijpers 2022). The Metapsy FAQ states that all data provided as part of the project are open access and that attribution is required. The source contains RCTs of psychotherapeutic interventions.

Data processing: We merged 20 individual datasets in Metapsy and performed some cleaning (for example, to extract study dates from titles). Almost all estimates are Hedges's g, but for one of the databases (`total-response`) we converted log-odds ratios to standardized mean differences (dividing by $\Pi/\sqrt{3}$ to scale SD of logistic regression) to keep a comparable scale. Signed z-values were computed as effect sizes divided by their standard errors.

Study characteristics:

- **Sample size and year:** `ss` is the sum of `n_arm1` and `n_arm2` when both are supplied; `year` is the source year or a year parsed from the study label.
- **Study ID:** we use the source `study` label within each database.
- **Meta-analysis ID:** we use the Metapsy database identifier.
- **Topic:** we use an explicit clinical-subject lookup; datasets `peer-support` and `total-response` are labeled as "mental health."
- Effects are standardized mean differences, mainly Hedges's g; a minority are converted from log odds ratios.

B.4. Arel-Bundock et al. 2022 (political science)

Reference: Arel-Bundock et al. (2022).

Research question: assess statistical power in political science research

Data availability: The replication package (a zip file dated `20241010`) is available at `https://osf.io/fgdet`. All data are in the public domain.

Data description and source: The authors searched for meta-analytic articles across 141 journals in political science and contacted their authors. The resulting source dataset contains 16,649 hypothesis tests, grouped in 351 meta-analyses and reported in 46 peer-reviewed meta-analytic articles.

Notes: This dataset collects meta-analysis inputs, rather than repeating data extraction from individual studies. It combines an earlier dataset from Doucouliagos et al. with another search. Therefore, some of this dataset will overlap with Askarov et al. However, the two articles aim to cover different fields and BEAR keeps the two datasets separate.

Data processing: We used the same code as the original authors (a replication package that uses makefiles), but skipped some additional processing they performed for their paper. BEAR uses `question_id` as `metaid` (351 groups defined by meta-analysis or research question). Signed z-values were computed as reported estimates divided by their standard errors. We verified that the distributions of z-values matched those in the article.

Study characteristics:

- **Study ID:** we use author's study ID; for **meta-analysis ID** we use `question_id`.
- `subset` column is the source `meta_id`
- We do not have information on the study design or measure of effect size.

B.5. Costello and Fox 2022; Yang et al. 2023, 2024 (ecology and evolution)

Reference: These two datasets come from the same replication package and were processed using the shared pipeline in Yang et al. 2024 (Yang, van Zwet, et al. 2024). The original references for the datasets are Yang et al. 2023 and Costello and Fox 2022 (Yang, Sánchez-Tójar, et al. 2023; Costello and Fox 2022).

Research question: decline effects in ecology (Costello and Fox) and publication bias and performance of empirical research (Yang et al)

Data collection: Costello and Fox examined 466 meta-analyses identified by searching Web of Science. The reproducibility dataset includes 88,000 rows across 466 analyses in 232 papers. Subsequently, Yang et al. removed duplicated studies and zeroes, discarding about 20% of rows. Yang et al. collected 102 meta-analyses published in relevant journals between 2010 and 2019, after removing records duplicated in the Costello and Fox dataset.

Data availability: The two datasets may be downloaded from `https://github.com/Yefeng0920/replication_EcoEvo_git/`. They are available under a CC BY 4.0 license.

Data processing: The only additional processing was to clean up study years. Signed z-values were computed as effect sizes divided by their standard errors. For Yang et al, BEAR uses the source file names as meta-analysis identifiers.

Costello and Fox:

In `BEAR.rds`, we relabel the effect measures as standardized mean differences, response ratio and correlation. The individual dataset keeps the original Yang et al. coding (Zr, lnRR, SMD).

Yang:

In `BEAR.rds`, we code Fisher's z as "correlation" and lnRR as "response_ratio" for consistency with other datasets.

Study characteristics:

- **Study ID:** for Costello and Fox, we use `study2`; for Yang, we use `study_ID`.
- **Meta-analysis ID:** for Costello and Fox, we use `meta.analysis.id`; for Yang, we use `meta_id`.
- **Topic:** for Costello and Fox, we use ecology, evolution or other ecology/evolution. Yang has no topic classification.
- **Costello and Fox:** the primary-study design is not classified.
- **Costello and Fox:** effects are standardized mean differences, log response ratios, Fisher's z correlations, or less common measures. Less common measures include regression coefficients, log odds ratios and mean differences. Some standardized mean differences are recorded as absolute values.
- **Yang:** the primary-study design is not classified.
- **Yang:** effects are standardized mean differences, log response ratios, or Fisher's z correlations.

B.6. What Works Clearinghouse (education)

Large database of education effect sizes created by the Institute of Education Sciences (Institute of Education Sciences 2025).

Snapshot date: 17 May 2025

Data availability: Flat files are available at `https://ies.ed.gov/ncee/wwc/studyfindings`. The Permissions and Disclaimers section of that website states: "Unless stated otherwise, all information on the U.S. Department of Education's IES website ... is in the public domain and may be reproduced, published, linked to, or otherwise used without permission from IES."

Data collection: The WWC source data have 13,054 findings (data rows) across 1,908 reviews. Each row is one finding, so a single source paper can contribute multiple rows. The dataset includes effect sizes, but not standard errors, as well as p-values reported by the studies and p-values re-calculated by the WWC reviewers. After BEAR processing and filtering, the main BEAR dataset includes 7,080 findings.

Data processing: We include only RCTs and quasi-experimental studies (98% of data altogether). Majority of these (about 80%) are RCTs.

We removed a few rows where the results were flagged as not meeting WWC standards or as ineligible for review. We also removed rows marked by WWC as subgroup measures to avoid treating subgroup analyses as independent focal findings. Where available, we preferred effect sizes and p-values calculated by WWC to the ones reported by the studies. We replaced p-values that were equal to zero (this occurred in 3% of the dataset) with a truncated p-value (i.e. assumed $p < 10^{-16}$). We calculated z-values from these p-values under the assumption of two-sided tests. We then applied the sign of the reported effect size, so z-values are signed where the effect size is non-zero.

Study characteristics:

- **Study ID:** we use citation-derived author and year; we use the original `WWCStudy` as fall-back in some cases.
- **Topic:** WWC dataset includes some useful topic classifications, but these are not included in `BEAR.rds` as they are either too granular or are overlapping flags for study categories

B.7. Cochrane Database of Systematic Reviews (medicine)

Trials with results available in CDSR, `https://www.cochranelibrary.com/` (Cochrane Collaboration 2025).

Snapshot date: 20 November 2025

Data collection: Using the R package `cochrane` (Schwab 2024), we combined data from 8,726 CDSR review records into a single file, `cdsr_interventions_19nov2025.csv` (63 MB).

Data processing: We applied several processing steps and extensive filtering.

During processing, we cleaned up study years and categorized measures (risk ratio, odds ratio, mean difference, etc.) and experimental designs. In particular, we created an "RCT" flag by scanning each review abstract for its inclusion criteria. Most studies in CDSR are RCTs, but some reviews also allowed quasi-experimental studies, so we could not classify many studies. We also classify outcome rows into `efficacy`, `safety`, `dropouts`, and `bias` based on the comparison, outcome, and subgroup labels.

To calculate effect sizes and z-values for BEAR, we used *metafor* to calculate standardized mean differences for continuous outcomes and probit-transformed differences for dichotomous outcomes. This puts continuous and binary outcomes on a comparable scale.

Unlike in most of the other datasets in BEAR, we filter CDSR data heavily. There are five important decisions:

1. Use only the outcome and comparison from each review coded as "1," as these are most likely to be the primary outcome and most relevant comparison.

2. Keep only rows classified as `efficacy`, excluding likely safety, dropout, and bias-related outcomes based on comparison, outcome, and subgroup labels.

3. Use only studies with continuous and dichotomous outcomes, removing rows where effect estimates are based on instrumental variables or individual patient data.

4. Remove rows where the effect measure is unknown (keeping OR, RR, Peto OR, mean difference, standardized mean difference, and risk difference) and rows with no participants.

5. Remove count data with no events or a 100% event rate in both arms (see the discussion of zero-event studies).

These processing steps are applied in the main dataset, `BEAR.rds`. Additional rows are still available to researchers in the individual dataset, `Cochrane.rds`.

As a result, the distributed `BEAR.rds` file includes a smaller analysis subset with about 36,000 rows: the first comparison and first outcome from each review, restricted to efficacy outcomes. The processed `data/Cochrane.rds` file is larger and is available for users who want to make different selection decisions. It contains 760,486 result rows, representing 90,042 review-study pairs across 6,619 Cochrane reviews.

A large proportion of CDSR rows have binary outcomes with no events in either arm or events in all participants in both arms. In the BEAR analysis subset before this exclusion, 2,371 of 26,619 binary rows (8.9%) met one of these conditions: 2,247 rows (8.4%) had no events in either arm and 124 rows (0.5%) had all participants experiencing the event in both arms.

Zero-event studies may be important for some metascientific analyses, but for modeling the distribution of z-values they are unhelpful. With these rows included, the empirical distribution has extra observations at or near zero, and the normal approximation is poor. Because of this peak at zero, including these non-significant studies has the counter-intuitive effect of increasing the

estimated degree of selection: the fitted distribution has a steeper drop-off between zero and the conventional significance threshold at 1.96.

Study characteristics:

- **Sample size and year:** `ss` is the sum of the two arm totals; `year` is `study.year`.
- **Study ID:** we use `study.name`.
- **Meta-analysis ID:** we use the Cochrane review identifier (`id`).
- **Topic:** we use the Cochrane specialty.
- `subset` records study-data provenance: published, unpublished, sought or mixed.
- `method`: We review meta-analysis eligibility criteria to determine which meta-analyses included only randomized trials. This classifies 69% of estimates as RCTs; the remaining 31% could not be classified reliably, although many are likely to be RCTs because Cochrane meta-analyses typically include RCT or quasi-randomized evidence.
- `measure`: About 32% of estimates are standardized mean differences and 68% are differences between probit-transformed event proportions.
- `topic` is Cochrane specialty and `subset` is source data type (published, unpublished, sought, or mixed); `outcome_group` is also available in the processed data.

B.8. Chavalarias et al. 2016 (PubMed/Medline)

Reference: Chavalarias et al. (2016).

Research question: reporting of p-values in the biomedical literature

Data collection: The authors used large-scale text mining of MEDLINE abstracts and PubMed full texts: 4.5 million p-values in 1.6 million MEDLINE abstracts, and 3.4 million p-values in 385,000 PubMed full-text articles.

Data availability: The extracted p-value dataset is publicly available at `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6FMTT3` (about 6 GB of data). According to the hosting platform, this dataset is released under the Creative Commons Attribution-NonCommercial 4.0 International License.

Data processing: Unsigned z-values were derived from p-values assuming two-sided tests. p-value truncation operators were processed by collapsing variants such as `<<`, `<<<`, `<=`, `less than`, and `=<` into `<`. We dropped 0.7% of rows where p-values did not have a "plain" format and 0.08% of rows where truncation could not be unambiguously classified. For the distributed version of BEAR 50,000 studies are selected at random, with one row per study, in order to keep the file size manageable.

Study characteristics:

- **Study ID:** we use the PubMed identifier (`PMID`).
- `subset` distinguishes estimates obtained by authors from MEDLINE abstracts vs from PMC full text.
- There is no information on study design (i.e. we cannot distinguish, for example, between prospective and retrospective studies or between RCTs and observational studies).
- There is no information on effect-size measure (for example, standardized mean difference vs. ratio).

B.9. SCORE (social science)

Reference: Tyner et al. (2026).

Research question: investigating the replicability of published claims in the social and behavioral sciences.

Data availability: The SCORE replication package is available at `https://osf.io/g5sny/`. The package includes replication data as well as reviews of claims made in the included papers.

Data description and source: BEAR includes two SCORE outputs. The source dataset of replications has 548 rows: 274 original claims and 274 matched replications. After BEAR processing and filtering, 267 matched replication rows are included. The source dataset for "all claims" has 3,066 rows for claims from the set of replicated papers (but no replications of these claims), where we select one statistic per claim; after BEAR processing and filtering, 1,942 claim rows are included.

Data processing: For the matched replication output, we used the package's converted correlation scale where available for original and replication statistics. For the set of all claims, effects are heterogeneous across the source papers.

When several inputs are available for constructing a z-value in the set of all claims, we prefer reported z, then reported t, then coefficient divided by standard error, then signed square-root F for numerator df 1, then a 95% confidence interval with a point estimate, then a two-sided p-value conversion. When the selected statistic only supplies a p-value and no sign can be inferred, the z-value is unsigned. In the text for all claims, it is typical to have multiple statistics backing up a single claim; for example, "F(1,38) = 3.73, p = .033, partial eta-squared = .16; F(1,39) = 7.28, p = .010, partial eta-squared = .16; F<1." We created a rule to pick one statistic per claim using agreement with the SCORE classification as significant or non-significant, the statistic's provenance, exactness, the available effect and standard error, and text order.

Study characteristics:

- **Study ID:** we use the DOI/source-paper identifier in both outputs.
- **Topic:** field of study (for example, "political science") as classified by the authors.
- The underlying study design is not systematically described.
- Measures: SCORE "claims" dataset measures include regression coefficients, standardized mean differences, correlations, and variance-explained statistics; but the effect measure is unclassified for over 80% of data. SCORE replications use correlations where available.

B.10. Lang 2025 (economics)

Reference: Lang (2025).

Research question: How frequently are statistically significant rejections of null hypotheses in recent empirical economics false positives?

Data collection: Lang samples empirical articles from the top five economics journals in 2021 and also articles previously collected in Brodeur, Cook, and Heyes, who sampled articles published in 2015 and 2018 in 25 leading economics journals. (See also the "Brodeur" dataset in BEAR for more details.) The sample consists mainly of natural, laboratory, or field experiments, excluding articles focused on theory, structural estimation, or econometric methods.

Lang re-extracted results from all articles rather than using the hypothesis tests extracted by Brodeur, Cook, and Heyes. Research assistants read the abstracts and results text and retained estimates emphasized in the abstract or described by the authors as principal results. When multiple specifications tested the same hypothesis, they retained the authors' preferred specification or the specification whose magnitude was used in subsequent analysis. Robustness checks and heterogeneity analyses were excluded unless the heterogeneity analysis was itself central to the article.

The source data contain 3,885 extracted hypothesis tests from 736 articles. Lang's main sample in the cited article restricts these to rejected null hypotheses with $1.96 \leq |t| < 10$, leaving 2,082 hypothesis tests from 663 articles. The dataset therefore does not identify one focal hypothesis per article. An article can contribute several hypothesis tests. However, only one preferred specification is retained for each hypothesis. (Lang paper also reports analyses restricted to one null hypothesis per article. The supplementary material contains a separate vector with one t-statistic for each of the 736 articles, but this vector does not contain row or article identifiers that allow it to be linked directly to the 3,885-test table. We therefore do not use this in BEAR data.)

Data processing: We use the 3,885 results. We use Lang's supplied, derounded t-statistic, derounded coefficient (`coef_dr` in the source data) and standard error (`se_dr`); the latter two are available in about 95% of cases. (Using either coefficients or t-statistics to calculate z-values leads to practically identical results; we use the latter.) We calculate the corresponding two-sided normal-approximation p-value. We retain some additional columns in `Lang.rds`.

We also pull DOIs for source articles using bibliographic metadata matching. These DOIs were obtained by us from Crossref, rather than supplied by Lang; the saved assignments are preserved when reprocessing the dataset. This is done automatically via Crossref. The matches are very good in almost all cases. For several edge cases we used LLM review to adjudicate.

Study characteristics:

- **Study ID:** `studyid` is `Lang_paper_` followed by Lang's `unique_paperid`, but with a few minor corrections from us (see GitHub repository for details). Article DOIs are also available as an additional column in `Lang.rds`.
- **Subset:** `subset` distinguishes observations drawn from the Brodeur–Cook–Heyes 2015/2018 source universe from Lang's 2021 top-five additions.
- **Method:** The `method` labels supplied with the Lang data mix study designs, identification strategies, and estimators. The large majority of estimates are "RCT," "DID," "IV," "RD," or "OLS." `Lang.rds` retains the original method labels. When constructing `BEAR.rds`, we collapse some label; for example, `RCT (DID)` and `RCT (IV)` become `RCT`; `DID (matching)` becomes `DID`; and `RD (DID)` and `RD (IV)` become `RD`. For estimates labeled `OLS`, `matching`, `SYSTEM-GMM`, or `IV-DID (matching)`, we retain the estimates but set `method` to missing.

- The dataset does not have data on type of measure to allow us to classify estimates as in some other datasets in BEAR (for example, risk vs. standardized mean difference).

B.11. Szucs and Ioannidis 2017 (cognitive neuroscience)

Reference: Szucs and Ioannidis (2017).

Research question: empirical assessment of published effect sizes, statistical significance, and power in the literature on cognitive neuroscience and psychology.

Data availability: The article and supporting data are available from PLOS Biology at `https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2000797`.

Data description and source: The source contains 26,841 text-mined t-test records from 3,801 papers in 18 journals (cognitive neuroscience, psychology, and medical). The processed dataset preserves all three fields; in the main BEAR dataset, we include only the cognitive neuroscience subset (since the psychology and medical literature are well covered by other datasets in BEAR).

Notes: The extraction targeted t-test records reported in article texts and did not mine tables.

Data processing: We followed our default procedure for deriving z-values from t-values.

Study characteristics:

- **Study ID:** to make look-ups of articles easier, we use `journal_<journal_code>_article_<article_id>`.
- **Topic:** we use the source `field` classification; note that `BEAR.rds` contains only cognitive neuroscience rows.
- The source data does not classify study designs.
- The original dataset labels all results as coming from a t-test, but there is no additional information on what was the underlying measure.

B.12. ClinicalTrials.gov snapshot (medicine)

Trials that reported results at ClinicalTrials.gov (National Library of Medicine 2025).

Snapshot date: 12 August 2025

Data source: This is a set of flat files of studies obtained from the AACT downloads page (`https://aact.ctti-clinicaltrials.org/downloads/`; over 2 GB of data for relevant tables), with relevant columns extracted and joined by us before BEAR processing.

Data availability: The terms and conditions for ClinicalTrials.gov are at `https://clinicaltrials.gov/about-site/terms-conditions`. We make distributions of derived effect sizes available in accordance with the terms and conditions. We only use publicly available data and do not imply any endorsement of this work by the National Library of Medicine, ClinicalTrials.gov or AACT.

Data processing: Processing was extensive. The full data table (`BEAR_data/clinicaltrialsgov.rds`) differs in some respects from the main BEAR dataset. We briefly discuss these differences here. Because of its rich structure, please refer to the data dictionary for the full data table.

Briefly, BEAR targets primary outcome results for completed studies using two data sources:

(1) Author-reported effect sizes

(2) Effect sizes derived from raw outcome data (group means, counts), alongside source data

Where both (1) and (2) are available, we prefer (1), but also make (2) available in the full data table. For (2), we use standardized mean differences for continuous outcomes and probit differences for count outcomes.

When working with the full dataset, variable `author_raw_overlap_key` can be used to link the two different calculations of the same effect; variable `include_in_bear` identifies rows included after applying this rule preferring author-reported estimates.

The main BEAR dataset keeps studies that (1) are randomized, (2) are interventional, and (3) have fewer than 20 candidate effect rows (see the description in this section).

The full ClinicalTrials.gov table uses only completed studies but does not apply these three restrictions. It contains 151,313 rows from 27,446 studies; 46.3% are author-reported and 53.7% are derived from raw data. The randomized interventional subset used in the main BEAR release contains 60,470 rows from 23,060 studies.

A single ClinicalTrials.gov study can contribute several primary outcome IDs, and each primary outcome can contain many measurements. The full data table records this as `n_effect_rows_per_study` and `n_outcomes_per_study`. The median study contributes one outcome ID and two rows to the full data table. Studies with 20 or more candidate effect rows remain in the full data table but are excluded from the main BEAR build.

In the full data table, some rows derived from raw data have `direction_unknown == TRUE`, meaning that the control/reference arm could not be reliably inferred, so the sign should be treated as random. Users interested in controlled trials or signed z-values should filter the results accordingly.

Appendix C explains how we created and chose the z-values based on p-values and confidence intervals. Briefly, we calculated z-values using two separate procedures (based on p-values and confidence intervals) and, if both were available, we made a judgment on which one was more appropriate using a predefined rule. ClinicalTrials.gov reports the number of sides for confidence intervals but not for p-values, so we treated p-values as two-sided. We dropped rows only when `z` was missing. Notably, infinite z-values can arise when `p_value` is recorded as exactly 0, and these are included because they are not missing.

Study characteristics:

- **Study ID:** we use the ClinicalTrials.gov trial identifier (`nct_id`).
- `subset` records trial phase.
- We include only randomized interventional studies, including parallel, crossover and factorial designs.
- About half of all effects (by which we mean both reported and our derived estimates) are standardized mean differences, mean differences, or probit differences. Another 14% are hazard, odds, risk, geometric, or rate ratios; 5% have the other category (in `BEAR.rds` this is coded as "other"), and 20% have no measure.

B.13. EU Clinical Trials Register (medicine)

Trials that reported results in the EU Clinical Trials Register, `https://www.clinicaltrialsregister.eu/` (European Medicines Agency 2025).

Snapshot date: 18 January 2026

Data collection: We used the R package `ctrdata` (Herold 2025) to create a snapshot of EU CTR (`collection = "euctr"`); we downloaded only trial records with results. We stored the downloaded records in a local SQLite database and extracted a fixed set of protocol and results fields plus several fields computed by `ctrdata` (trial phase, sample size, and a "first primary endpoint" p-value/size derived by the package). We use only the EU CTR portion of this extraction in BEAR.

Data availability: EUCTR is publicly accessible. The EUCTR legal notice notes that publication on EUCTR does not constitute an endorsement of the information reported.

Data processing: Processing was extensive. We use only primary endpoints. We default missing CI levels to 95% and where truncation of p-values was not stated we assume equality (the latter is the case for about 15% of data). Where the data do not report whether a test is one- or two-sided, we assume two-sided. The rest of the derivation of z-values follows the standard procedure for dealing with p-values and confidence intervals described in Appendix C. We set `year` to the year of entry since completion dates were unavailable. We dropped rows only when `z` was missing; infinite z-values are included. We harmonize phase labels to follow the same naming convention as ClinicalTrials.gov.

Study characteristics:

- **Sample size:** `ss` is the extracted endpoint sample size (`n`).
- **Study ID:** we use the EudraCT trial identifier (`id`).
- `subset` records trial phase.
- The registry extract does not provide a row-level randomisation indicator, so BEAR does not classify the trial design.
- Effect measures are derived from reported estimand labels; about half of these are mean differences, hazard ratios, odds ratios, risk differences, or risk ratios. Another one third have other category (in `BEAR.rds` this is coded as "other"), and 23% have no measure.
- `subset` records registry phase.

B.14. Barnett and Wren 2019 (medicine)

Reference: Barnett and Wren (2019).

Research question: Bias for statistical significance in health and medical journals

Data availability: The dataset `Georgescu.Wren.RData` may be downloaded from `https://github.com/agbarnett/intervals/tree/master/data` or `https://github.com/jdwren/ASEC`. The repositories give no clear license, but the BMJ Open article is CC BY 4.0.

Data description and source: The dataset is a collection of confidence intervals for ratio estimates from Medline papers published between 1976 and 2019. Authors scraped (via regular expressions, followed by an independent check using another data mining algorithm, manual checks for 10,000 intervals) 968,000 CIs from abstracts and 350,000 from full texts.

Notes: These are "ratio estimates" such as odds ratios, hazard ratios and risk ratios. Binary outcomes tend to contain (much) less information than continuous outcomes.

Data processing: We exclude a small fraction of data. First, we restrict the data to 95% confidence intervals: where `ci.level` was missing we assume it was 0.95; among intervals with a known `ci.level`, about 0.3% were not 95% and are dropped. We drop intervals with non-positive width (i.e. `lower >= upper`). We use the log scale (we replace zero or negative lower bounds with a small positive constant; this affects about 0.1% of the sample) and back out an approximate standard error and point estimate under the usual normal approximation for a 95% CI, setting the point estimate to the midpoint of the interval.

For the main version of BEAR dataset, `BEAR.rds`, 50,000 studies are selected at random, with one row per study, in order to keep the file size manageable.

Study characteristics:

- **Study ID:** we use the PubMed identifier (`pubmed`).
- There is no information on study design (i.e. we cannot distinguish, for example, between prospective and retrospective studies or between RCTs and observational studies).
- Effects are ratios on a logarithmic scale, without distinguishing odds, risks, hazards etc.

B.15. Open Science Collaboration 2015 (psychology)

Reference: Open Science Collaboration (2015).

Research question: replications of a quasi-random sample of 100 experiments in psychology.

Data collection: The authors make a comprehensive dataset of results available with their paper. BEAR includes only the replications because all original studies in that set have $p < 0.05$.

Data availability: The data file `rpp_data.csv` is available in the paper's repository, `https://osf.io/ytpuq/overview`, under a CC0 1.0 Universal license.

Data processing: We derived z-values from the replication p-values assuming two-sided p-values (we use the p-value column for the replication, not the p-value for the original study). We applied the sign of the replication correlation, so z-values are signed where that correlation is available. We treated p-values recorded as exactly 0 as truncated and recorded this as a lower bound on z (z-operator `>`); otherwise we treated p-values as exact (z-operator `=`). We retained the replication sample size as `ss`.

In `BEAR.rds`, the measure is recorded as "correlation," because that is the column from original dataset that we use.

Study characteristics:

- BEAR treats the replication studies as RCTs.
- The retained effect measure is the replication correlation.

C. Standard procedure for dealing with p-values and confidence intervals

Most datasets in BEAR come with pre-computed effect sizes and standard errors. Some, especially databases of clinical trials and large sets of scraped studies, report p-values. Moreover, reporting in clinical trial databases is done by investigators, and the data-entry fields are not standardized, requiring more extensive clean-up and a procedure for deriving z-values. Lastly, in some cases datasets contain confidence intervals and effect sizes, but z- and p-values may be missing.

C.1. Data cleaning and default values

For confidence interval levels, we set implausible values (such as those above 100 or below 0) to missing and standardize proportions and percentages: investigators sometimes report 90 and sometimes 0.9 for a 90% confidence interval. In EU CTR, confidence levels can be missing, and we assume 95% by default. Some data state whether CIs or p-values are one- or two-sided.

Similarly, when working with p-values we assume a two-sided test when no additional information is stated.

C.2. Confidence intervals

When working with confidence intervals, we analyze ratio measures on the log scale when the lower bound, upper bound, and point estimate are all strictly positive; we keep non-ratio measures on the raw scale.

To derive standard errors from confidence intervals, we first compute the critical value. To avoid undefined critical values at 100% or extremely close to it, we bound the implied α at a small value. We then compute two one-sided standard error estimates from the upper and lower bounds and use their average when both are finite.

We compute a simple symmetry diagnostic for the confidence interval on the chosen scale, defined as the smaller of the two half-widths divided by the larger. We treat intervals as "symmetric enough" only if this ratio exceeds 0.8. This arbitrary threshold affects when we use the CI-derived z-statistic; otherwise, we use a p-derived z-statistic where one is available.

Where additional information is available, we look for potentially non-Wald intervals by matching text in the parameter-type column. We search for keywords such as median, Hodges, posterior/Bayes, exact, Fieller, bootstrap, and permutation. For flagged estimates, we prefer to derive the z-value from the p-value.

C.3. P-values

When a source reports t-statistics and degrees of freedom, we first derive the two-sided p-value from the t distribution and then derive the z-value (retaining the sign of the original t-statistic, if available).

For clinical trials databases (ClinicalTrials.gov and EU CTR), we sometimes find ambiguity in how p-values are recorded. Some significant results report p-values close to 1, suggesting that authors report the cumulative distribution, $\Phi(z)$. Therefore, for each row we also calculate $2 \cdot \min\{p, 1-p\}$. When a CI-derived z-value is available, we choose the p-value that gives the closest z-value.

To avoid dropping rows, we retain unsigned z-statistics when the sign cannot be determined. If the sign of the effect is known, we apply it to the p-derived z; otherwise, we use its absolute value.

C.4. Choice of a single z-value when multiple candidates are available

In situations where z can be derived from both CIs and p-values, we choose a single "best" z-statistic per row as follows. We prefer the CI-derived z when it looks like a plausible Wald interval on the chosen scale, meaning: (1) finite z and standard error, (2) positive standard error, (3) no non-Wald flag from a keyword search of the measure label, and (4) either missing symmetry information or symmetry above 0.8. Otherwise, we use the p-derived z.

C.5. Truncation and infinite z-values

For reported two-sided p-value bounds, $p < c$ gives a lower bound on $|z|$. BEAR records this inequality using `z_operator`. The operator applies to the magnitude, including when the stored z-value is negative. For example, a stored value of -1.96 with operator `>` means a negative direction and magnitude greater than 1.96, not $z > -1.96$.

Exact zero p-values can produce infinite z-values. We retain infinities in BEAR source data, although signal-to-noise models truncate these infinite values.