Decline effects, statistical artifacts, and a meta-analytic paradox

Richard D. Morey[1]

[1] Cardiff University

Author Note

All code to reproduce this paper can be found at https://github.com/richarddmorey/decline_bias. This draft was compiled Monday 09 Feb 2026 at 11:24:40 GMT.

The authors made the following contributions. Richard D. Morey: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Richard D. Morey, 70 Park Place. E-mail: moreyr@cardiff.ac.uk

Abstract

The decline effect (Protzko & Schooler, 2017) is an observed phenomenon where effect sizes in experiments apparently diminish in size from the first paper demonstrating the effect to later replications. This has been taken as a symptom of an unhealthy scientific ecosystem, possibly caused by the "winner's curse" (selection on significance and regression to the mean), publication bias or opportunistic analyses. I show that decline effects can arise as an artifact from a much simpler source: the original article determining the sign of the effect in a meta-analysis. Moreover, such artifactual decline effects will show correlations with some of the same experimental properties that one would expect from biases from poor behavior, such as the sample size of the original study.

Keywords: decline effect

Word count: X

Decline effects, statistical artifacts, and a meta-analytic paradox

Introduction

Because scientists often cannot directly observe a whole system, they often make inferences from phenomena that are deemed signatures of some underlying cause: morphological similarities for descent from a common ancestor (Darwin, 1859), cosmic microwave background radiation for the Big Bang (Dicke, Peebles, Roll, & Wilkinson, 1965), or additivity in response times for serial cognitive processing (Sternberg, 1998). Signatures are consequences of particular ways that a system might work.

A weakness of using signatures, however, is that purported signatures may be observed for reasons other than the putative cause. The worst possible way this could go wrong is if there is an *artifactual* reason for the signature: one due merely to the manner in which a phenomenon is studied.

Statistical metascience is built on signatures that are used to infer issues in scientific behaviour. For instance, the observation that effect sizes appear to diminish over time after an initial discovery—called the "decline effect" (Protzko & Schooler, 2017)—is taken as a signature of poor scientific behavior (either individual or systemic). Pietschnig, Siegel, Eder, and Gittler (2019) say that

> "We show in the present meta-meta-analysis evidence for overproportional. . . and stronger effect declines than increases in the published intelligence literature. Effect misestimations are most likely due to low initial study power and strategic research and submission behaviors of exploratory researchers and replicators alike." (p. 12)

They suggest reforms meant to ameliorate the problems they infer.

I will show that such signatures can arise as an artifact of the meta-analytic study of such effects. Seemingly paradoxically, I will show that biased estimates of meta-analytic

quantities can arise from unbiased estimates of effect sizes, without any poor behaviour (assuming the most basic statistical model for the outcomes, every study published, and no opportunistic reporting). I will then suggest a resolution of the meta-analytical paradox, proposing that it arises from the stripping of away of the research context in a way that is central to the meta-scientific perspective.

Meta-analytic "coining"

"Coining" is the practice of aligning effect sizes by their observed for interpretability. For instance, Ioannidis (2008) (Figure 2, p. 65) demonstrated how selection for significance can lead to a relationship between effect size and sample size. He first aligned all observed effect sizes so they were positive.

Coining is also used in the study of decline effects. In this context, the sign of the effect of an initial study (the earliest one in the set, often the first to report the effect in question) is flipped to be positive, and then all subsequent studies are re-aligned to be consistent with that direction (e.g. Fanelli, Costas, & Ioannidis, 2017; Gong & Jiao, 2019; Nuijten, van Assen, Augusteijn, Crompvoets, & Wicherts, 2020; Pietschnig et al., 2019). For instance, Fanelli et al. (2017) report in their methods that:

> "Each dataset was standardized following previously established protocols (14). Moreover we inverted the sign of (i.e., multiplied by -1) all primary studies within meta-analyses whose pooled summary estimates were smaller than zero—a process known as 'coining.'" (p. 3719)

Coining appears to be standard practice in metascience, and may not be explicitly called "coining". For instance, Open Science Collaboration (2015) reported that

> "To be able to compare and analyze correlations across study-pairs, the original study's effect size was coded as positive; the replication study's effect size was

> coded as negative if the replication study's effect was opposite to that of the original study." (p. aac4716-3)

and reported a "substantial decline" (p. 943) in effect sizes from original to replication (see also e.g. Camerer et al., 2018; Errington et al., 2021).

In the study of decline effects, coining is used to yield an common definition of decline across studies: a "decline" is when the coined effect size becomes less positive. Otherwise, a decline of a positive effect (e.g. moving from 1 to .5) might be cancelled by decline of a negative effect (e.g. moving from -1 to -.5). Coining aligns the signs across all studies so that many research areas can be studied simultaneously, drawing far-reaching conclusions across science.

Data-dependent reporting practices

Consider any research context in which there two possible outcomes that can be supported, and these outcomes are symmetric in the sense that neither outcome is inherently favored from a reporting perspective. Gender or sex differences are a common example; one could report a result as favoring men or women.

There are different ways to choose how to report a difference in such contexts. If the research is confirmatory or following up on previous studies that showed a similar effect, one might choose the to align a pre-registration or a previously reported effect. For instance, if we predicted that women would score higher on a particular inventory than men, then when the opposite is observed we might report a negative effect size. Likewise, if we have a theoretical reason to expect one direction or another, we might choose to report from the perspective of the theory.

But what about novel effects? Another choice is simply to report the effect from the perspective of the group that scored the highest, i.e. that makes the difference positive. Johfre and Freese (2021) explicitly recommend "relying on the values of the [differences] to

make a decision [on the reference category]. Given that positive numbers are cognitively simpler than negative values, the reference category can be chosen such that the presented coefficients are positive." Given that this will have no effect on inferential statistics, a data-driven approach is a reasonable choice.

I call reporting practices based on outcomes "data-driven reporting practices" (DRPs). Note that these are not the same as hypothesizing after the results are known (indeed, they may be agnostic to any hypothesis) or cherry picking. Data-driven reports are perfectly transparent with respect to the data; they simply choose a *reporting perspective* based on the outcome. There is nothing questionable about such reporting *per se*.

The practice of coining in meta-analysis is equivalent to re-aligning effect sizes as if the original authors had used DRP, and then subsequent authors had reported their effect sizes in line with the original study. As a shorthand for both this particular DRP and the meta-analytic practice of coining, I will use the term "sign alignment". For the purpose of this paper, it does not matter *who* does it; I will explore its effects on meta-analysis.

Statistical assumptions. For shorthand, call the experimental context A and the two outcomes/perspectives a_1 and a_2 (e.g., A might be "response time in a visual search task" and a_1 might be "women/women were faster on average" and a_2 "men/men were faster on average"). Of course, there is a difference between the perspective and the outcome; which I am referring to will be clear from context.

Any experimental outcome X_i ($i = 1$ an initial study, and $i = 2, \ldots$ for replications) is an unbiased estimate of estimate of some underlying true effect size μ with normal error:

$$X_i \sim Normal(\mu, \sigma_i)$$

where σ_i is the standard error or estimate X_i. To begin, we assume that there will be one replication, that the initial and replication outcomes are statistically independent of one another, that both the initial study and the replication will be reported, and have the same true mean. Arbitrarily, we assume that μ is the effect size when a_1 is taken as the research

perspective (correspondingly, $-\mu$ is the effect size when a_2 is taken as the research perspective).

We assume that all results are "coined" after being observed. Equivalently, 1) all initial studies report their findings by stating the effect size from the perspective that was favored by the data, and 2) replications will adopt the perspective of the initial study. The initial effect size reported will thus always be positive.

Let us call the coined result Y_i, which can be thought of as a pair of numbers.[1] For reports from initial studies Y_1,

$$Y_1 = \begin{cases} (|X_1|, a_1) & \text{if} X_1 > 0 \\ (|X_1|, a_2) & \text{if} X_1 < 0 \end{cases}$$

and for replications (where $i > 1$),

$$Y_i = \begin{cases} (X_2 sgn(X_1), a_1) & \text{if} X_1 > 0, i > 1 \\ (X_2 sgn(X_1), a_2) & \text{if} X_1 < 0, i > 1 \end{cases}$$

where $sgn\, x$ is the sign function that returns -1 when $x < 0$ and 1 when $x > 0$ (ignoring for simplicity the zero probability event that an experimental result is exactly 0).

Whenever X_1 and X_2 disagree in sign, the effect sizes in Y_1 and Y_2 will also disagree in sign. The difference between the results X_1, X_2 and reported results Y_1, Y_2 are that Y_1, Y_2 have been "aligned" to the outcome of the initial study.

It is obvious that we can obtain an unbiased estimate of μ by reversing the sign of the effect size report in Y_i when a_2 is reported, then taking the average with all the reports where a_1 is reported. An estimate of $-\mu$ (the effect from the opposite perspective) could found in the same way. There is nothing defective about the DRP from the perspective of

[1] In a slight abuse of notation, when I refer to Y_i by itself in mathematical formulae, assume that it is shorthand for the first number: the reported, sign-aligned effect size.

estimating the underlying effect size, as long as one takes care to understand the perspective that each study takes.

Bias in decline effects

We now switch to a meta-scientific perspective. Suppose we ask the question: *do effect sizes tend to decrease from the initial to the replication?* Assume that for every initial report Y_1, we also have a report from a replication Y_2. In order to assess whether reported effects decline, we compare the numerical effect size in the report of Y_1 ($|X_1|$) with the numerical effect size in the report of Y_2 ($X_2 sgn(X_1)$) to produce a "decline effect" d_s (where the s is for "sign" to indicate the effect has been aligned to the sign of the initial study):

$$\begin{aligned} d_s &= Y_1 - Y_2 \\ &= |X_1| - X_2 sgn\, X_1 \end{aligned}$$

If $d_s > 0$, this is taken to mean that the effect size has "declined" to some extent from the initial study. Although I frame this section as comparing an initial study to a replication, my critique here applies to any comparison between two kinds of studies (e.g. observational vs. randomized control trials, Franklin et al., 2017).

When both X_1 and X_2 have the same true mean μ—that is, there is no true decline effect—will the expected decline effect d_s be equal to 0? Surprisingly, no: The expected value of d_s will *always* be larger than 0 under these circumstances. There is always an artifactual bias in the decline effect in precisely the direction that metascientists use as a signature.

To see why, consider the situation where the true effect size μ is 0. The signs of the initial and replication will differ in 1/2 of cases. Anytime this occurs, Y_1 is positive and Y_2 is negative; hence, a decline is observed. When the signs agree (probability 1/2), the

probability that $Y_1 > Y_2$ is 1/2 (because they have the same mean). Thus, the probability of observing a decline in this scenario is $1/2 + 1/4 = 3/4$, despite there being no decline effect.

Figure 1A depicts the situation graphically for two true effect sizes. The figure shows the patterns of results that would lead to the identification of a decline effect. As can be seen, most of the space is occupied by "decline". When the true effect size is 0 and there is no decline ("a"), 3/4 of the bivariate distribution would count as a "decline". When the true effect size is larger ("b"), only about half the bivariate distribution would be in the "decline" region.

The key insight is that the sign alignment happens in one direction or the other with some probability: whether the result needs to be "coined" is itself random. Unless we have a specific *perspective* pinned down before the meta-analysis (e.g., a_1 or a_2), the perspective will thus be random.

When there is no decline effect, the expected value of d_s is $COV(X_1, sgn\, X_1)$ (proof in appendix), which will be a function of how far, in standard error units, μ is from 0. The amount of bias is shown in Figure 1A, as a function of the true effect size in standard errors of X_1. The larger the true effect (in either direction), the less probability of the observed sign differing from the true sign; hence the source of bias diminishes.

Bias in funnel plots, Egger's tests, and PET

Unlike in decline effect analysis, meta-analysts do not generally use coining. Coining is used across research context to align all the effect sizes; funnel plots and related tests (e.g. Egger's regression test and the precision-effect test, PET) are about looking at a single research context, so explicit sign alignment is not necessary. However, as I point out, sign-alignment can also be done by the original authors as a data-dependent reporting practice. It turns out that sign alignment will bias also funnel plots and the related tests.

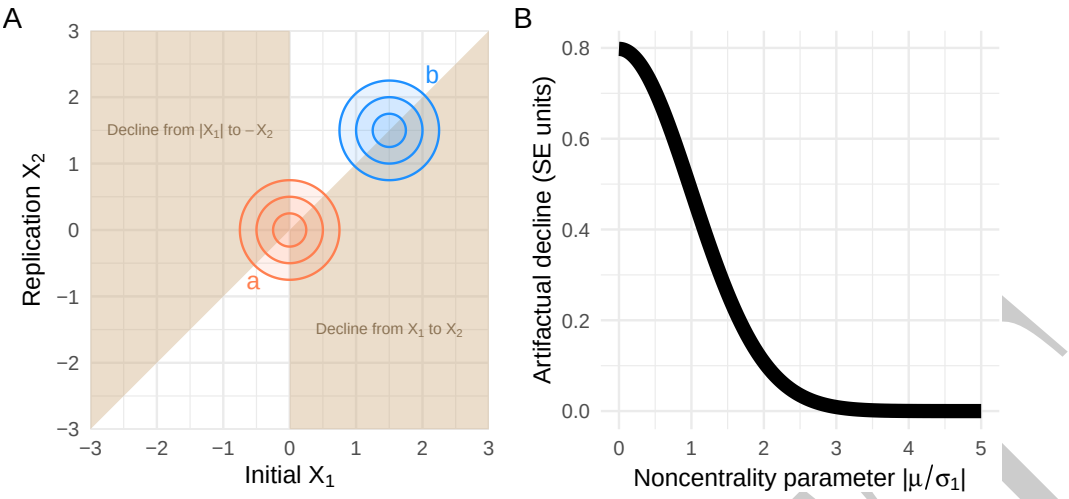

Figure 1. A: Result combinations that would lead to a conclusion of "decline", if the error-prone sign of the first experiment determines the interpretation of the effect (shaded regions). Labels "a" and "b" denote two hypothetical bivariate effect size distributions with true effect sizes 0 and 1.5, respectively. See text for explanation. B: Amount of bias in the decline effect as a function of the noncentrality parameter μ/σ_1 of the initial study.

To show why sign alignment introduces bias into funnel plots, consider a simple situation in which we perform an initial experiment and a replication. Both experiments have the same true effect size ($\mu = 0.20$). The initial experiment has a standard error of 0.20 for estimating μ, while the replication has 4 times the sample size, yielding a smaller standard error of 0.10. The initial experiment has a 16% of being observed in the wrong direction.

Consider what happens without sign alignment, but *conditional* on the sign of the initial experiment. Figure 2A depicts the situation graphically. Each experiment/sign possibility is shown as a distribution. The points show the expectation of the observed effect size conditional on the sign of the initial experiment. The sizes of the distributions and points are proportional to the probability of that outcome.

Conditional on a positive initial outcome, the initial effect size is slightly overestimated on average. This is exactly counteracted by the underestimation when the

initial outcome is negative, weighted by the lower probability of the negative outcome. With no sign alignment, the sign of the initial experiment has no effect on how the replication is reported, hence both replications have the same conditional mean: exactly μ.

Drawing the conditional meta-regression lines (dotted) up to the $y = 0$, we can see that the two lines are biased in opposite directions. When the initial effect is positive, the slope is slightly negative; when the initial effect is negative, the slope is strongly positive. The *average* meta-regression line, however (solid, thick blue line), is exactly vertical and intersects with the x-axis at exactly $x = \mu$. This is the expected behaviour for a funnel plot.

Figure 2B shows the situation with sign alignment. Sign alignment does two things: when the initial experiment has a negative sign, it is flipped to be positive; second, replication outcome signs are also flipped to be consistent this interpretation. When the observed effect in the initial experiment is in the correct direction—positive—nothing changes: the conditional meta-regression line is the same. But when the initial experiment has a negative outcome, the sign alignment flips the conditional meta-regression line across $y = 0$.

Both conditional regression lines point in the same direction and they cannot balance one another out. The average meta-regression line (solid, thick red line) has an x-intercept that does not correspond to the true effect size. Moreoever, on average it has a negative slope, which is supposed to be a signature of poor scientific behavior.

As the number of studies in the set increases, the bias in the intercept does not disappear, though the bias in the slope tends to 0 if there is no publication bias. The source of the bias is the sign alignment of the initial study, and hence all subsequent studies will be affected in the same way. As with the bias in decline effects, the bias in funnel plots due to sign alignment decreases for larger effect sizes because the probability of an initial sign flip gets smaller; see proof in Appendix.

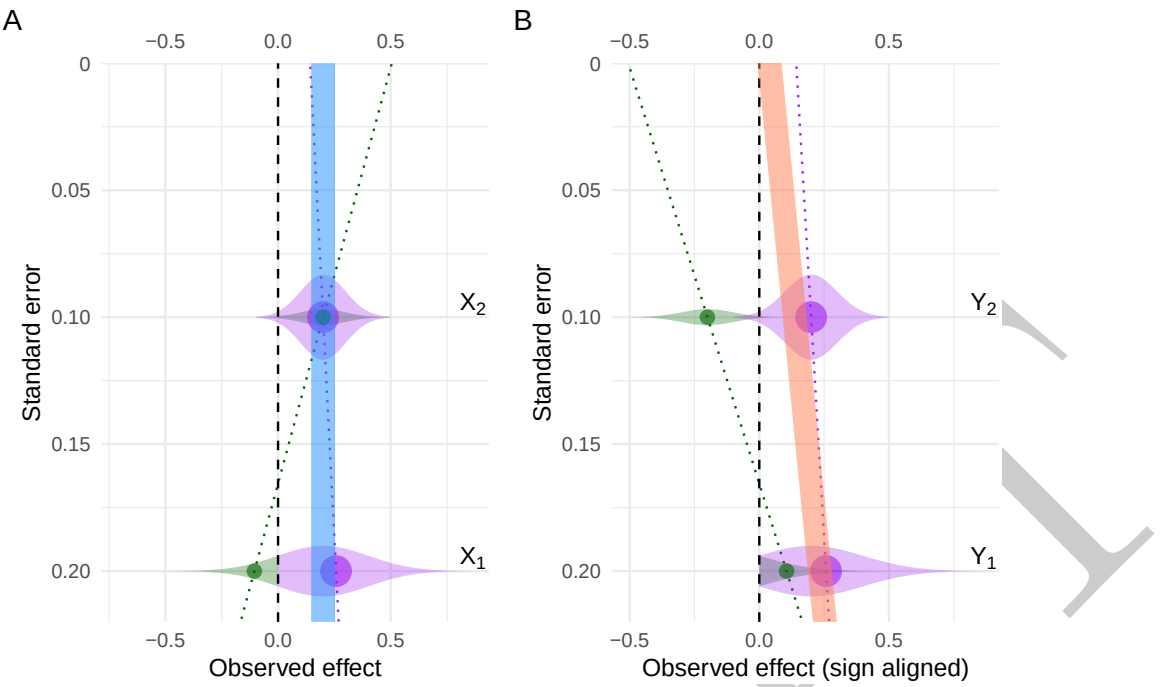

Figure 2. Demonstration of the source of sign-alignment bias in funnel plots. See text for explanation.

Simulation

In this section we continue using our simple two-experiment (initial/replication) setup, but we will vary the sample sizes and true effect sizes. I give details about the simulation setup in the Appendix; here, I give an abbreviated version that is enough to understand the simulations without burdening the reader with too much formal information.

Each simulated experiment is assumed to have two independent groups whose mean difference will be the effect size of interest. For simplicity, the true variance of each group is known to be 1; we can thus treat the mean difference as a standardized effect size. In the initial study, the sample size in the two groups are equal and drawn from the distribution shown in Figure 3A. For the replication, the sample sizes in the groups were assumed to be 4 times larger than in the initial study. The true effect sizes for each pair of initial study and replication were assumed equal, drawn from the distribution shown in Figure 3B.

For each of the 2 (initial/replication) $\times$5000 simulated studies, an observed difference (X) was drawn assuming normal error. These differences were then coined to produce sign-aligned effect sizes (Y). The simulation's original/replication simulation setup is similar to the setup of Open Science Collaboration (2015).

In the simulations that follow, for each statistic I show that there is no bias if X is considered; however, using Y in place of X will artifactually yield meta-scientific signatures of poor scientific behaviour.

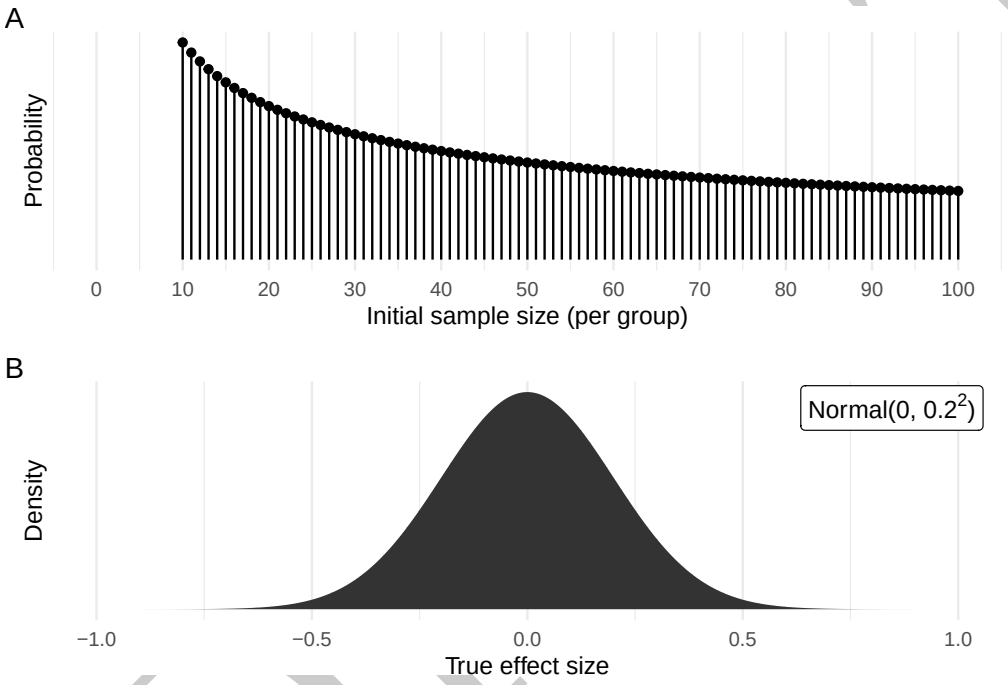

Figure 3. Distributions from which initial per-group sample sizes (A) and true effect sizes (B) were drawn for the reported simulations. All samples were independent.

Relationship between effect size and sample size

A relationship between effect size and sample size is often cited as evidence for publication bias; for example, Stanley, Carter, and Doucouliagos (2018) cited such a correlation as part of the "quite clear" evidence for publication bias, because "inverse correlation between the magnitude of the effect size and sample size would be expected

when there is selective reporting for statistical significance" (p. 1328). Ioannidis (2008) purported to demonstrate the effects of publication bias by showing a relationship between effect size and total sample size in a set of meta-analyses. Likewise, the logic of funnel plots and meta-regression techniques such as Egger's regression and PET-PEESE depend on finding such relationships.

Figure 4A shows the relationship between the initial effect size X_{1j} and the sample size in the 5000 simulations. By design, these were sampled to be independent, so there is no relationship between them. Figure 4B, however, shows that once the results are sign-aligned, the relationship appears. Although Ioannidis (2008) intended to show that selection on statistical significance would show this effect, he would have seen the same effect even had he not selected on significance simply due to his explicit coining of the effect sizes.

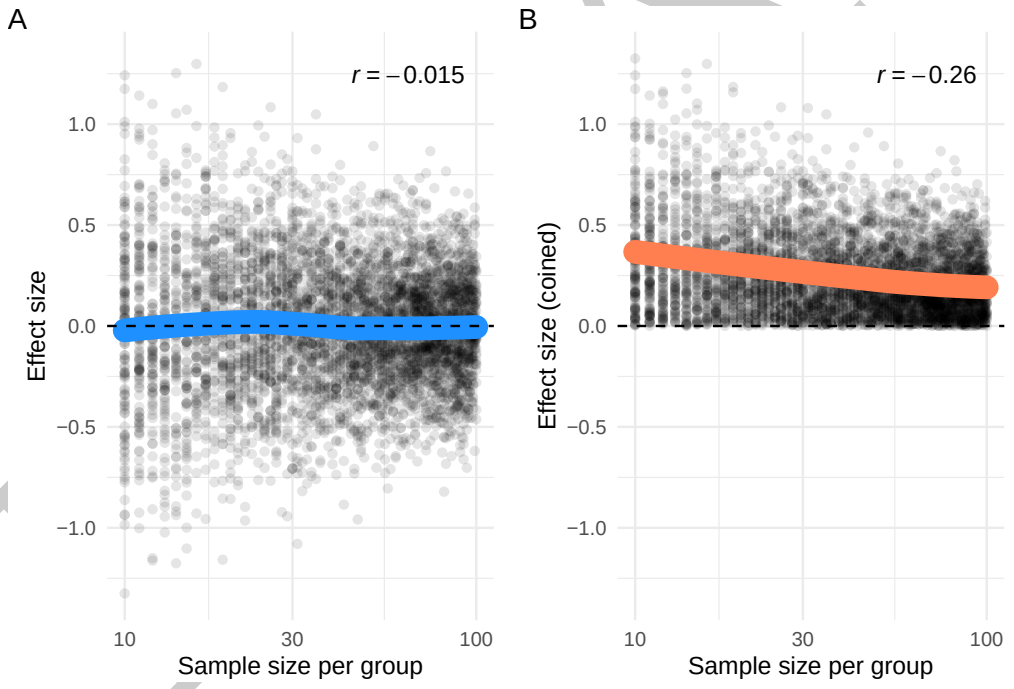

Figure 4. Relationship between the initial observed effect sizes and initial sample sizes for (A) non-sign-aligned results and (B) sign-aligned results. Each point is a single simulated initial result. The thick line in each plot shows the LOESS nonparametric regression curve.

Decline effects

To demonstrate the effect of sign-alignment on estimates of the decline effect, I computed the difference between the initial and replication estimates $X_{1j} - X_{2j}$ (unaligned) or $Y_{1j} - Y_{2j}$ (sign-aligned).

Figure 5A shows the unaligned differences. Because these simulated results have the same mean, as expected, the average difference is 0. However, as Figure 5B shows, when the results are sign aligned, the results show artifactual decline effects. Although there is no true decline effect in any simulation, the apparent decline effect is large: on average, the effect sizes "decline" by over 50%.

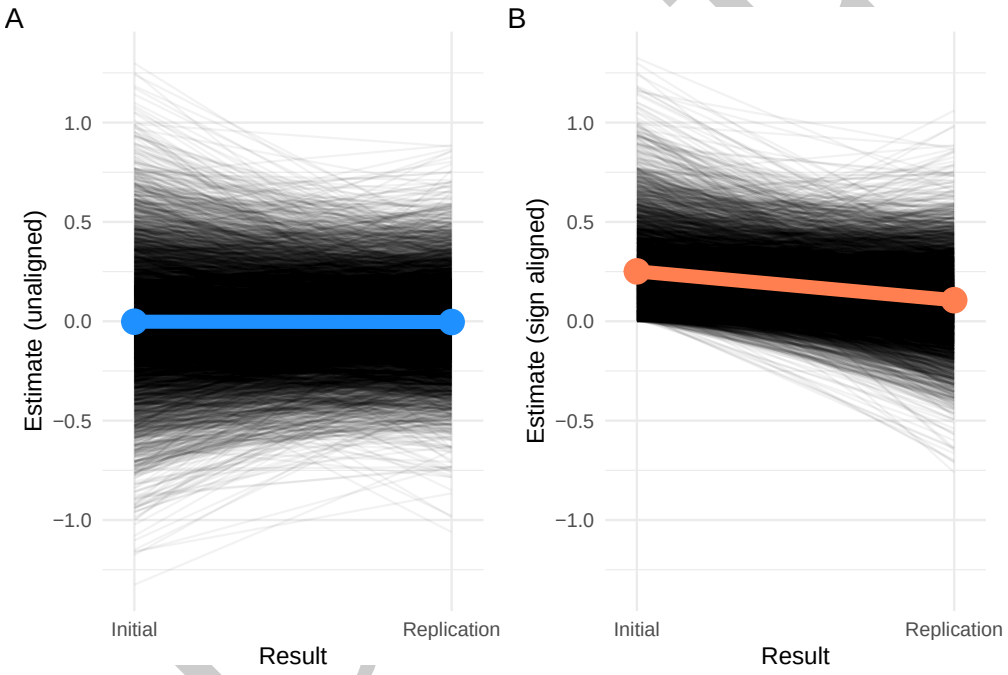

Figure 5. Initial and replication estimates without sign-alignment (A; X_{1j} and X_{2j}) and with sign-alignment (B; Y_{1j} and Y_{2j}). Each thin black line shows a single simulation. The thick line in each graph shows the average.

Correlation between decline effect and sample size. Meta-analysts look for relationships between decline effects and other properties of initial papers to understand "risk factors" for bias in literatures. For instance, Pietschnig et al. (2019) report that

"[e]ffect misestimations were more substantial when initial studies had smaller sample sizes and reported larger effects, thus indicating suboptimal initial study power as the main driver of effect misestimations in initial studies" (p. 1).

Relationships between decline effects either initial effect sizes or sample sizes can be explained artifactually. A relationship between initial effect size and both unaligned or aligned decline is explainable by simple regression to the mean. Figure 6A and B show this relationship in the simulated data. We replicate the strong relationship Pietschnig et al. (2019) with initial effect size without any difference between initial and replication true effect sizes. This has relationship has nothing to do with "suboptimal initial study power"; it is a mere statistical artifact.

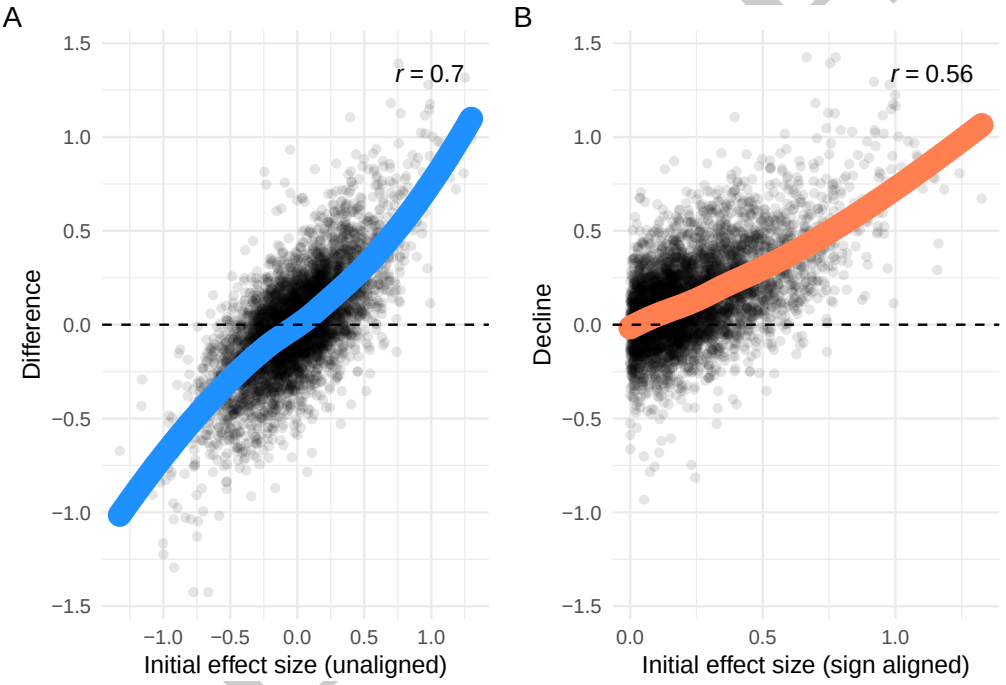

Figure 6. Correlation between initial effect size and the difference/decline from initial to replication study in unaligned (A) and sign-aligned (B) results. Each point is the result of a single simulation. The thick line in each plot is a LOESS nonparametric regression curve.

Relationships between initial sample size and the decline effect can be equally attributed to a statistical artifact, but the cause is somewhat subtler. As Figure 7A shows,

no such relationship is apparent in the simulated data before sign-alignment; indeed, the data were simulated in such a way that these quantities were independent. However, Figure 7B shows that after sign alignment, the relationship hypothesized and reported by Pietschnig et al. (2019) appears. This artifactual relationship can be attributed to the fact that that the bias in computing the decline effect is a decreasing function of the noncentrality parameter (the effect size in standard error units), and the noncentrality parameter is larger when N is larger.

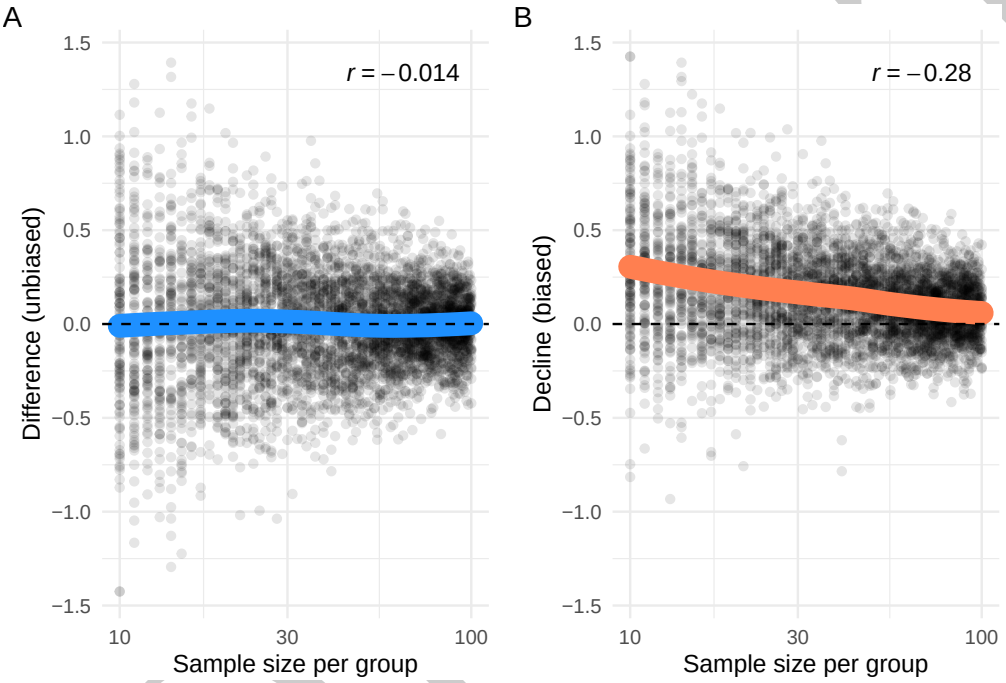

Figure 7. Correlation between sample size and the difference/decline from initial to replication study in unaligned (A) and sign-aligned (B) results. Each point is the result of a single simulation. The thick line in each plot is a LOESS nonparametric regression curve.

Asymmetric funnel plots

As previously demonstrated, sign-alignment is expected to have the effect of biasing funnel plots to 1) be asymmetric, and 2) have x-intercepts away from the true effect size even in the absence of publication bias.

For the purposes of simulating funnel plots, I added a second simulated replication (X_{3j}) of the same size as the first replication to prevent the meta-regressions from being trivial (only consisting of two points). For visualization purposes, instead of showing the individual points, I show the meta-regression line for each simulation, along with the average meta-regression line.

Meta-regression lines were built from least squares estimates for the model:

$$X_{ij} = b_{0j} + b_{1j} s_{ij}$$

where s_{ij} is the standard error for the ith study ($i = 1, 2, 3$) in the jth simulation. Sign-aligned meta-regressions were obtained through corresponding meta-regressions for Y_{ij}. For the purposes of this simple example, weighting the points by precision is unnecessary.

Figure 8A shows the meta-regression lines of standard error onto observed effect size for the unaligned data points ($X_{ij}, j = 1, 2, 3$). On average, the funnels are symmetric and have an x-intercept of 0, which is the average effect size in the data.

Figure 8B shows the corresponding meta-regression lines for the sign-aligned data. The lines are, on average, skewed to have negative slopes. The average x-intercept is not 0, but is instead less than 0. As is clear from the previous discussion (particularly Figure 2), this x-intercept is a function of the true average effect size and the standard errors of the replications; it could be shifted right or left by changing the relationship between the initial and replication standard errors.

We can also assess the bias over all simulations in the x-intercept as an estimate of the true effect size. Figure 9A shows the relationship between the true effect size and error in the meta-regression estimate of the effect size for the unaligned results. For the range of true effect sizes, the meta-regression estimate appears to be unbiased (though quite variable, because we only used three studies).

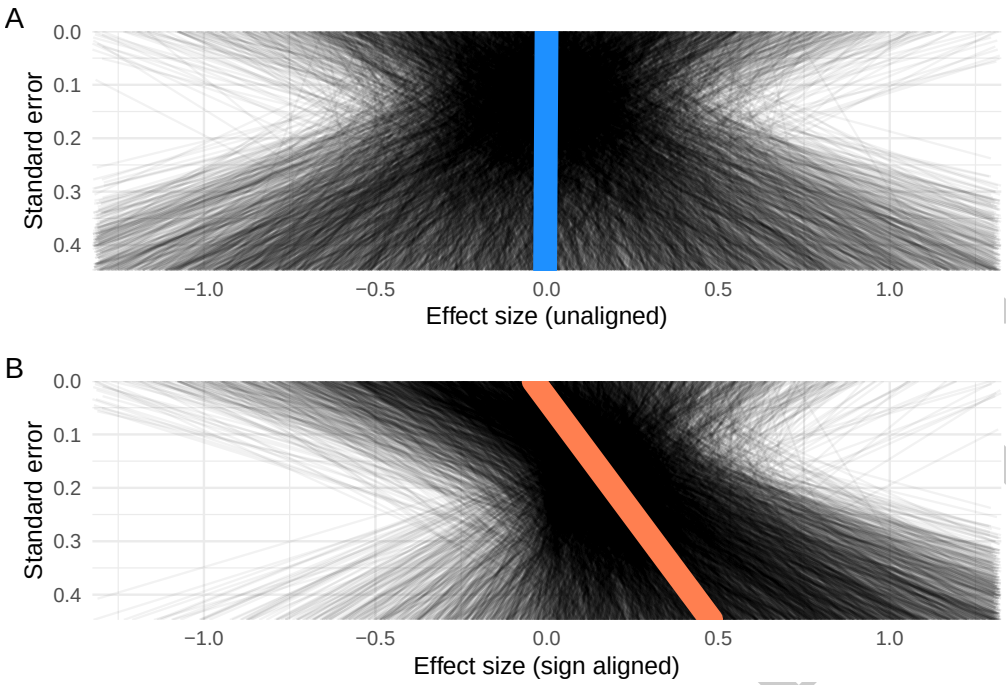

Figure 8. Meta-regression lines for each of the simulations using the non-sign-aligned data (A) or sign-aligned data (B).

The same is not true of the sign-aligned estimates. After sign-correcting the true effect sizes (i.e. when the sign is flipped, we account for the fact that we are estimating $-\mu$), the sign-corrected estimates are not only biased on average, as shown in Figure 8B; Figure 9B shows that bias appears to be an increasing function of the true effect size.

Finally, I note that all of the sign-alignment bias noted above is due to the inclusion of the initial sign-aligned study in the meta-analysis itself. If we drop the initial study from the meta-analysis and consider only the first replication (for instance), the replications are unbiased estimates of the true effect size (once the true effect sizes have been suitably sign-corrected as well).

Figure 10 shows the relationship between the true effect size and observed, sign-aligned replication effect size. The bias caused by the initial study and sign-alignment has disappeared. This reveals that a data-dependent reporting strategy itself is not a problem; the problem is meta-analysts assuming that the effect sizes of the initial studies

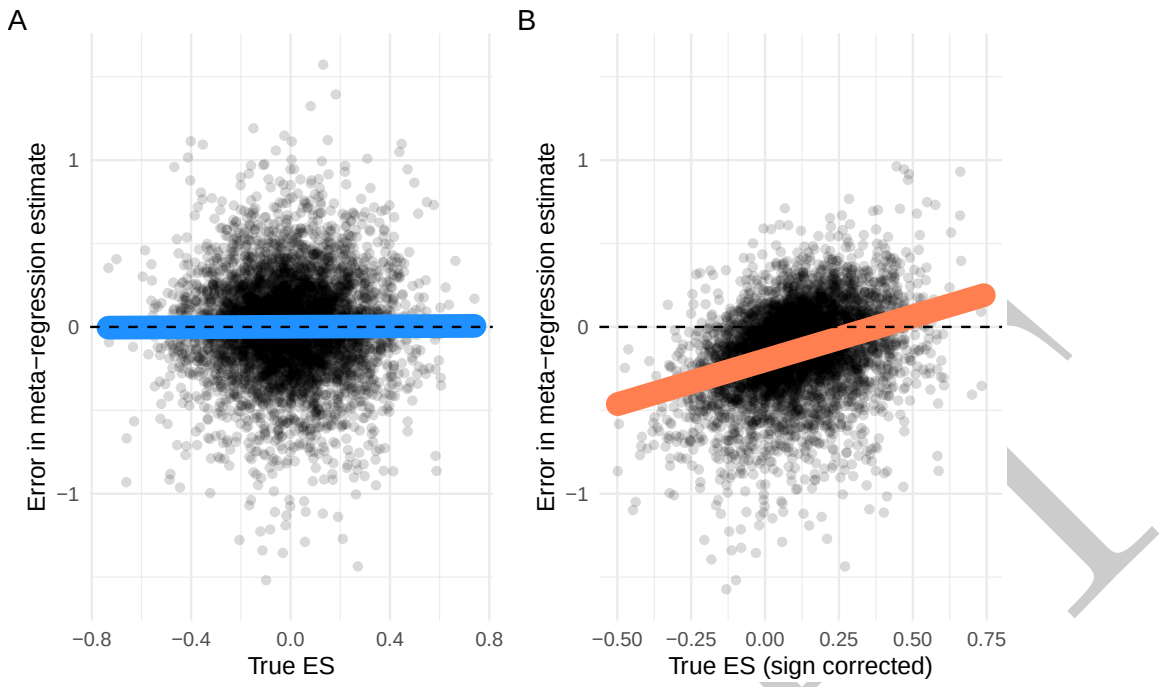

Figure 9. Error in the meta-regression estimate of the effect size in the simulations (intercept of the regression of effect size onto standard error, minus the true effect size) as a function of true effet size in non-sign-aligned simulations (A) and sign-aligned simulations (B). In B, the true effect size is corrected when the sign of the initial study was flipped (i.e. the true effect size is $-\mu$ in these cases).

can be modeled in the same way as later replications, when it fact this may not be true.

A meta-analytic paradox

I have shown that sign-alignment leads to analysis artifacts of the sort that one would expect under poor scientific behavior. This is clearly undesirable. "Coining" by meta-analysts and sign-alignment as a DRP are equivalent, yet I have called the DRP "innocuous". Is this contradictory? Also, how can reports by original authors that I am claiming lead to biased differences be themselves unbiased? At first glance this appears to be a paradox.

The resolution of the paradox must lie in the different goals of those reporting the

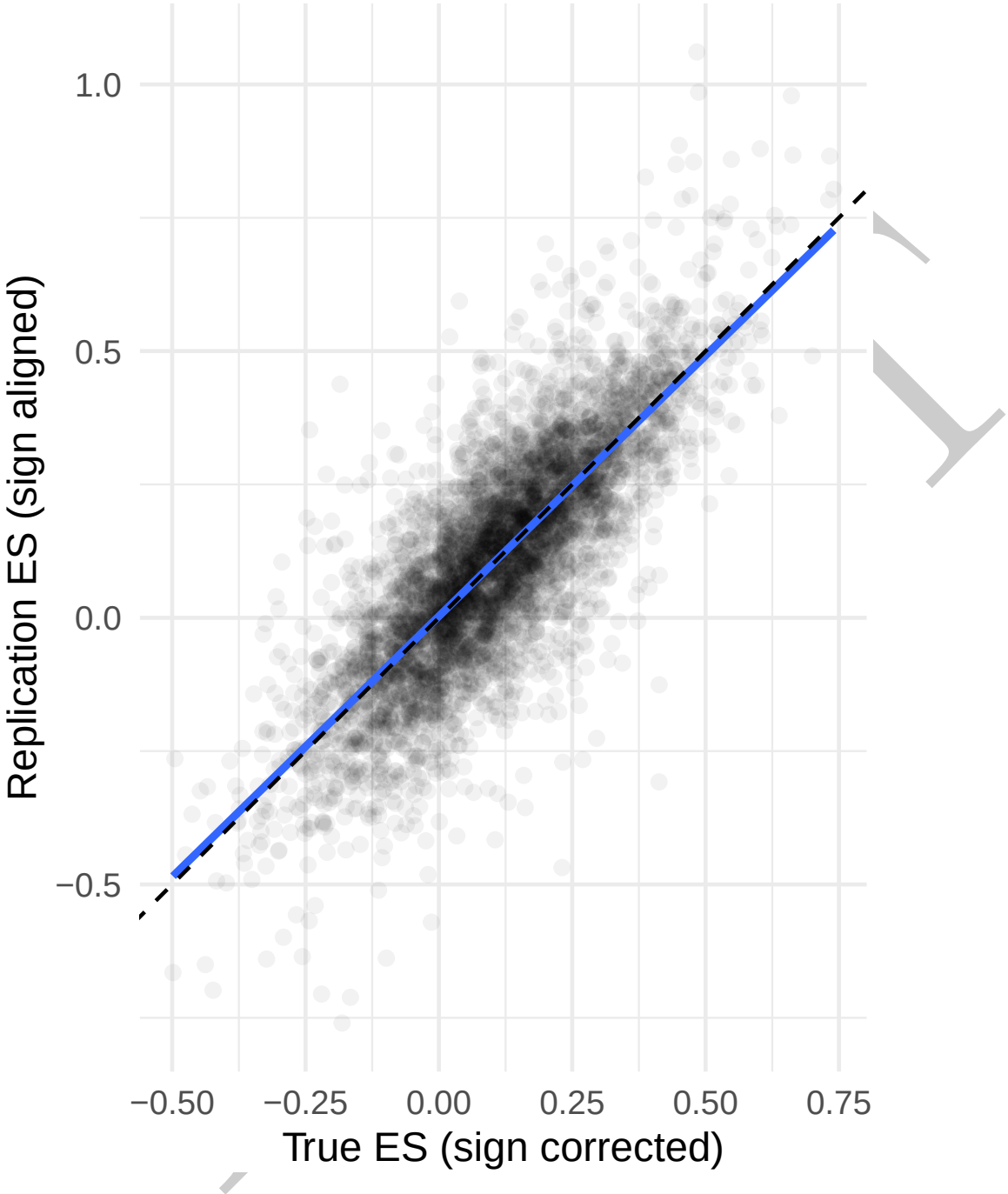

Figure 10. Relationship between the observed, sign-aligned replication effect size and the true sign-aligned effect size for all simulations. The solid blue line is the least-squares regression line; the dashed black line is the line $y = x$. The true effect size is corrected when the sign of the initial study was flipped (i.e. the true effect size is $-\mu$ in these cases).

statistical results and those meta-analysing them. A person reporting an effect size in a novel experiment has reported the result transparently and accurately even if they preferentially report it in the positive direction. They have met the goals that someone reporting results should have.

The meta-analyst, on the other hand, has a different goal: to combine results across studies to come to a well-supported conclusion. By coining—or by assuming that the results have not been sign-aligned, when in fact they have been—they have undermined that goal. Their conclusions will be biased.

The mistake that the meta-analyst has made is in assuming something that is not true about the data: that sets of results can be treated as sequences of independent random variables with symmetric error. Sign-alignment—and perhaps other DRPs—violates that assumption. The results are not independent because the data from the first study was used to perform an operation on all studies.

But there is a way around the bias induced by sign-alignment: the effect size in an original report must be interpreted more holistically. In the notation introduced earlier, when sign alignment happens, the effect size is a actually a *pair* of values $(|X_1|, a_1)$ - the first value denoting the effect size, and the second the perspective from which it is reported. An unbiased estimate of the effect size from either perspective can be obtained by aligning all effects on a_1 or a_2. The perspective by which one aligns must be independent of the data (e.g., from a previous study, a theory, a pre-registration, a Bayesian prior, or a design asymmetry). Without a fixed perspective, the perspective becomes random and this is where the trouble begins.

Discussion

The bias due to coining in meta-analyses was previously pointed out by Franklin et al. (2017; see also commentary by Sterne, 2018), but does not appear to be widely known.

Currently, Franklin et al. (2017)'s excellent commentary has fewer than twenty citations according to Scopus, and meta-analysts continue to use the technique. Morrissey (2016) discusses naïve meta-analyses of absolute values and other nonlinear transformations of data and points out that meta-analytic "findings" may be mere artifacts of poor statistical models. Here I show that coining introduces these kinds of artifacts, and DRPs may introduce them as well.

Although the problem with coining (and related data transforms) have been noted before in other literatures, the point that the innocuous DRPs may cause equivalent effects appears to be novel. This raises the general possibility that the simple assumptions underlying most meta-analyses—that effect size reports can be treated as independent observations with symmetric error around a fixed true value—may often be false in important ways even when there is no problematic behaviour among scientists.

How much does this effect any particular meta-analysis? This is unknown, and I am not arguing here that this artifact accounts for any particular previous finding. Other issues (e.g. publication bias) can also cause the effects, and not all research contexts will be as susceptible to sign alignment. I do, however, take it for granted that meta-analysts should not use methods *known* to be biased. If a meta-analyst wants to draw an inference in any research context, the burden of evidence is squarely on them to argue that their inference is not plagued by artifacts such as those from sign-alignment or other DRPs. The use and interpretation of methods that may be sensitive to the issues I describe either implicitly or explicitly assumes that they are *not* a problem.

It has been argued that many experiments in psychological science have effect sizes that are small enough to lead to a high probability that a replication may invert the sign merely by chance (see e.g. Open Science Collaboration (2015), which reported that 17 of 97 replications showed an opposite sign to the initial study). Small effects with reports depending on the data will lead to artifacts when studying decline effects. Interestingly, the

reformers using the biased methods in meta-analyses are the ones that argue that the conditions exist that would lead to maximum bias.

Recommendations

Meta-analysis should never coin effect sizes. I believe it should go without saying that meta-analysts should not introduce a source of bias into their meta-analyses, even if it appears to improve interpretability. Meta-analysts must find a way outside the data itself to align effect sizes. In some cases this will be straightforward (e.g., treatment/control designs or based on theory); in others, it will be less straightforward. In any case, however, it seems that meta-analysts cannot avoid engaging with the research contexts of the results they study.

Meta-analysts should not include the initial study in most meta-analyses. If it is uncertain whether an initial report was affected by DRPs, it is simply safer to exclude it. The simple act of excluding the initial study will entirely mitigate the effect I have pointed out, because it is caused by the error in the initial study. Exclusion of the initial study would also mitigate regression to the mean (sometimes called the "winner's curse," Young, Ioannidis, & Al-Ubaydli, 2008), which seems unavoidable otherwise.

Meta-analysts should seek to understand the role of DRPs in meta-analysis. As I have shown, reporting practices that have no effect on the inferences in individual studies may nevertheless have an effect on meta-analyses. If the the simple practice of reporting a novel effect in the direction in which it is observed can bias meta-analyses—and this has remained unexplored for decades—this raises the possibility that other DRPs may effect meta-analyses.

As I have previously argued (Morey & Davis-Stober, 2025), it is crucial for the health of science that meta-scientific tools should have good properties (e.g. not find evidence for poor behaviour) when scientific behaviour is not problematic. We do not have to agree on

every aspect of good or poor scientific behavior to agree that methods yielding artifactual decline effects should not be used. If meta-analysts continue to use biased methods, the rhetoric used to attack poor-quality science now can (and will) be used against good-quality science in the future.

Appendix

Proofs

Bias in decline effect. Let X_1 and X_2 be two independent random variables; the mean and variance of X_1 will be denoted μ_1 and σ_1^2, and likewise for X_2. X_1 and X_2 are assumed to be effect size estimates from initial and a replication experiment.

Let $s(x)$ denote the sign function $sgn\, x$. Then the decline effect is a random variable D_s defined as:

$$D_s = s(X_1)X_2 - s(X_1)X_2$$

Let s_1 be $E[s(X_1)] = 2p - 1$ where $p = Pr(X_1 > 0)$. The expectation of D_s is

$$\begin{aligned} E[D_s] &= E[s(X_1)X_1 - s(X_1)X_2] \\ &= E[s(X_1)X_1] - E[s(X_1)X_2] \\ &= E[s(X_1)X_1] - s_1\mu_2 \\ &= COV[s(X_1), X_1] + s_1\mu_1 - s_1\mu_2 \\ &= COV[s(X_1), X_1] + s_1(\mu_1 - \mu_2) \end{aligned}$$

Under the assumption of no decline effect ($\mu_1 - \mu_2 = 0$), $E[D_s] = COV[s(X_1), X_1]$.[2] Of course, $s(X_1)$ and X_1 will be positively correlated, so the decline effect estimate will be biased, in general (for all $SE < \infty$).

When there *is* a difference between μ_1 and μ_2, and assuming X_1 has a normal distribution (so $s_1 = 2\Phi(-\mu_1) - 1$ where Φ is the CDF of the normal distribution), the bias in the estimate of the difference D_s relative to $\mu_1 - \mu_2$ will be

$$COV[s(X_1), X_1] - 2(\mu_1 - \mu_2)\Phi(\mu_1)$$

However, one may object to this definition of "bias" when $\mu_1 \neq \mu_2$ because it does not capture the logic of "attenuation" (that is, if both effect sizes are negative when $\mu_1 - \mu_2$ is positive it does not represent an attenuation; it is an increase).

Another option that better captures attenuation might be; e.g.,

$$d_a = |X_1| - |X_2|$$

as a measure of decline. However, this estimator, too, will suffer from bias because $E(|X|) > |E(X)|$ when X can be negative and there is variability in X. The more variability, the greater bias; hence, if X_2 is a more precise estimator than X_1—which would often be the case if X_2 arises from a replication or a meta-analysis—the decline effect will again be overestimated, and will not be 0 when the true decline effect is 0. One may also object to this definition of decline because extreme sign differences may not be picked up as problematic (i.e. from -1 to 1 is a large change, but the "decline" as the difference in absolute values is 0).

[2] If one is unwilling to assume independence of X_1 and X_2 (e.g. if X_2 is a meta-analytic effect estimate including X_1), the bias when there is truly no decline effect will be $COV[s(X_1), X_1] - COV[s(X_1), X_2]$. The first term will dominate when X_1 is a small component of X_2.

Any useful definition of "decline" will likely be more complicated than simple arithmetic operations, and may depend on the scientific context. Regardless of how decline is defined, I propose the following necessary condition: If a pair of published results (original/replication) would be defined as a decline, then the reversed pair (treating the replication as the original, and vice versa) should *not* indicate a decline. Sign-alignment can violate this simple rule.

Bias in funnel plot meta-regression. Let $K \geq 2$ be the total number of points in the meta-regression. We assume that the first study (X_1) is the one that was used for sign alignment, and our vector of observed effect sizes is

$$\mathbf{Y} = [Y_1, Y_2, \ldots, Y_K]'.$$

The expected value of $\mathbf{Y}$ is

$$E(\mathbf{Y}) = [E(|X_1|), (2p-1)\mu \mathbf{1}'_{K-1}]'.$$

Let e_i be the standard error of study i and $S = \sum_i (e_i - \bar{e})^2$. Applying the least squares solution the expected slope b_1 and intercept b_0 are:

$$\begin{aligned} E(b_1) &= \frac{e_1 - \bar{e}}{S} \left(E(|X_1|) - (2p-1)\mu \right) \\ E(b_0) &= qE(|X_1|) + (1-q)(2p-1)\mu \end{aligned}$$

where $q = 1/K - \bar{e}(e_1 - \bar{e})/S)$. If we assume normality so that

$$E(|X_1|) = \frac{\exp\{-\mu^2/2\}}{\sqrt{\pi/2}} + (2p-1)\mu,$$

we obtain

$$\begin{aligned} E(b_1) &= \frac{e_1 - \bar{e}}{S\sqrt{\pi/2}} \exp\left\{-\mu^2/2\right\} \\ E(b_0) &= \frac{q \exp\{-\mu^2/2\}}{\sqrt{\pi/2}} + (2p-1)\mu \end{aligned}$$

It is obvious that as $\mu \to \infty$ and thus $p \to 1$, $E(b_1) \to 0$ and $E(b_0) - \mu \to 0$. Likewise when $\mu \to -\infty$ and thus $p \to 0$, $E(b_1) \to 0$ and $E(b_0) - (-\mu) \to 0$ (i.e., $E(b_0)$ approaches the effect size with flipped sign).

Bacause $S \to \infty$ as $K \to \infty$, $E(b_1) \to 0$ and $E(b_0) \to (2p-1)\mu$ also as $K \to \infty$. Thus the bias in the slope diminishes, but not the bias in the intercept.

Simulation details

Let N_{ij} be the sample size per group for the ith study in the jth experimental context ($i = 1$ for the initial study, $i > 1$ for replications). We sampled the initial sample sizes N_{1j}, then based the replication sample sizes on these:

$$\begin{aligned} \sqrt{N_{1j}} &\overset{indep.}{\sim} \text{Uniform}(\sqrt{10 - 1/2}, \sqrt{100 + 1/2}), \\ N_{ij} &= 4N_{1j},\ i > 1, \end{aligned}$$

and N_{1j} was rounded to the nearest integer. The quantity $\sqrt{N_{1j}}$ was sampled from a uniform distribution so that, once squared, small sample sizes would be more common than larger ones. Replication sample sizes were assumed to be 4 times larger than the initial studies, though this does not matter much because the bias in the decline effect is only a function of the initial study properties.

The standardized effect size μ_j for all studies in an experimental context was assumed to be the same:

$$\mu_j \overset{indep.}{\sim} \text{Normal}(0, 0.20^2).$$

A mean effect size of 0 was chosen to respect the symmetry in the assumption that either group could act as a reference.

The effect size X_{ij} before sign alignment was then sampled from a Normal with mean

μ_j and standard error $\sqrt{2/N_{ij}}$:

$$X_{ij} \overset{indep.}{\sim} \text{Normal}(\mu_j, 2/N_{ij}).$$

Sign-aligned effect sizes were then computed:

$$Y_{ij} = \begin{cases} |X_{ij}| & i = 1 \\ X_{ij} sgn\, X_{1j} & i > 1 \end{cases}$$

References

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. John Murray.

Dicke, R. H., Peebles, P. J. E., Roll, P. G., & Wilkinson, D. T. (1965). Cosmic Black-Body Radiation. *The Astrophysical Journal*, *142*, 414–419. https://doi.org/10.1086/148306

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, *114*(14), 3714–3719. https://doi.org/10.1073/pnas.1618569114

Franklin, J. M., Dejene, S., Huybrechts, K. F., Wang, S. V., Kulldorff, M., & Rothman, K. J. (2017). A Bias in the Evaluation of Bias Comparing Randomized Trials with Nonexperimental Studies. *Epidemiologic Methods*, *6*(1), 20160018. https://doi.org/10.1515/em-2016-0018

Gong, Z., & Jiao, X. (2019). Are Effect Sizes in Emotional Intelligence Field Declining? A Meta-Meta Analysis. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01655

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, *19*(5), 640–648. Retrieved from https://www.jstor.org/stable/25662607

Johfre, S. S., & Freese, J. (2021). Reconsidering the Reference Category. *Sociological Methodology*, *51*(2), 253–269. https://doi.org/10.1177/0081175020982632

Morey, R. D., & Davis-Stober, C. P. (2025). On the Poor Statistical Properties of the P-Curve Meta-Analytic Procedure. *Journal of the American Statistical Association*,

0(0), 1–13. https://doi.org/10.1080/01621459.2025.2544397

Morrissey, M. B. (2016). Meta-analysis of magnitudes, differences and variation in evolutionary parameters. *Journal of Evolutionary Biology*, *29*(10), 1882–1904. https://doi.org/10.1111/jeb.12950

Nuijten, M. B., van Assen, M. A. L. M., Augusteijn, H. E. M., Crompvoets, E. A. V., & Wicherts, J. M. (2020). Effect Sizes, Power, and Biases in Intelligence Research: A Meta-Meta-Analysis. *Journal of Intelligence*, *8*(4), 36. https://doi.org/10.3390/jintelligence8040036

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6521), 943.

Pietschnig, J., Siegel, M., Eder, J. S. N., & Gittler, G. (2019). Effect Declines Are Systematic, Strong, and Ubiquitous: A Meta-Meta-Analysis of the Decline Effect in Intelligence Research. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02874

Protzko, J., & Schooler, J. W. (2017). Decline effects: Types, mechanisms, and personal reflections. In *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 85–107). Hoboken, NJ, US: Wiley Blackwell. https://doi.org/10.1002/9781119095910.ch6

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Sternberg, S. (1998). *Discovering Mental Processing Stages: The Method of Additive Factors*. https://doi.org/10.7551/mitpress/3967.003.0017

Sterne, J. (2018). Commentary: Does the selective inversion approach demonstrate bias in the results of studies using routinely collected data? *BMJ*, *362*, k3259. https://doi.org/10.1136/bmj.k3259

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why Current Publication

538 Practices May Distort Science. *PLOS Medicine*, *5*(10), e201.

539 https://doi.org/10.1371/journal.pmed.0050201