Appendix C

Appendix C

This supplementary material describes the underlying data generation process of the simulation.

1 Original study generation

To generate original experimental studies, the study-level true effect size θ_k was drawn from normal distributions $N(\theta, \tau^2)$, where θ denotes the overall population effect (0, 0.2, and 0.5 in the form of standardized mean difference) and τ the between-study standard deviation (0.020, 0.050, and 0.125). The within-study standard deviation σ was independently sampled from $U(0.5, 2.5)$ for each study, allowing variability in outcomes across studies. Participant-level data of the original studies were generated for two independent groups, a treatment group and a control group. The true population mean of the control group (μ_c) was sampled from a standard normal distribution, $\mu_c \sim N(0, 1)$, representing a baseline outcome with no treatment effect. The true population mean of the treatment group was defined as $\mu_t = \mu_c + \theta_k \sigma$. The outcome variable was then sampled from normal distributions $Y_c \sim N(\mu_c, \sigma^2)$ in the control group and $Y_t \sim N(\mu_t, \sigma^2)$ in the treatment group.

The *p*-hacking and publication bias mechanisms worked jointly as the biasing mechanism to determine which original studies were “published” for replication in the replication study phase. Each simulation cycle began by generating a single original study under one of three *p*-hacking levels: no, medium, or high *p*-hacking. The likelihood of using specific *p*-hacking practices was determined by the predefined proportions described earlier. In the unbiased condition, the simulation generated participant-level data of two independent groups of equal size and computed the observed standardized mean difference (Hedge’s *g*), its sampling variance, and standard error. Under biased conditions, simulation created two correlated dependent variables per participant using a bivariate normal distribution with correlation $r = 0.20$. The simulation then implemented combinations of outlier exclusion, selective reporting of dependent variables, and optional stopping on the resulting data, which were then passed to the publication bias mechanism. This mechanism probabilistically decided whether the study would be “published” based on its statistical significance and effect direction.

This process was repeated iteratively: new studies were generated, subjected to *p*-hacking, and filtered through the publication bias mechanism until the target number of 500 published studies was reached. As a result, the final pool of original studies represents the visible portion of a biased research literature: one shaped simultaneously by questionable research practices and selective publication.

2 Replication study generation

Participant-level data were generated based on the same true study-level true effects (θ_k) obtained from the original studies. Each replication study consisted of two independent groups, including a treatment group and a control group of equal size. For each replication, the true mean of the control group (μ_c) was sampled from a standard normal distribution, $\mu_c \sim N(0,1)$, representing a baseline outcome with no treatment effect. The true mean of the treatment group was defined as $\mu_t = \mu_c + \theta_i \sigma$, where σ denotes the within-study standard deviation, fixed at 1 for all replications to standardize outcome variability across studies. Given these parameters, each participant's outcome was drawn from $N(\mu_c, \sigma^2)$ in the control group and $N(\mu_t, \sigma^2)$ in the treatment group. For each replication, the standardized mean difference (Hedge's *g*) and its associated variance and standard error were computed. The sampling distribution of the observed replication effect size therefore reflects variability due to both sampling error and the study-level true effect θ_k inherited from the corresponding original study. For each original true study effect θ_k, sets of $k = 2,5,10$ replications were simulated, with per-group sample sizes of $n = 40,100,400$, respectively. Each of these replication scenarios was repeated 500 times to stabilize the estimates