Supporting Information

Fanelli et al. 10.1073/pnas.1618569114

SI Methods

During December 2013, we searched Thompson Reuters' Web of Science database, using the string ("meta-analy*" OR "meta-analy*" OR "meta analy*") as topic (which captures terms used anywhere in the database record, including, title, abstract, keyword, keywords expanded), restricting the search to document types "article" or "review." This search strategy was identified as the most efficient way to retrieve suitable and usable meta-analyses. Sampling was stratified by scientific discipline, by restricting each search to the specification of journal names included in each of the 22 disciplinary categories used by Thompson Reuters' Essential Science Indicators database. These disciplines and the abbreviations used throughout the text and figures are defined in *Methods* in the main text.

All potentially relevant records were retrieved for each discipline and their order was randomized so that subsequent phases of inspection and inclusion would yield randomized samples. From each discipline, an initial list of potentially relevant records was retrieved based on title and abstract. The pdf file for all these records was retrieved. When the pdf was not available, attempts were made to contact the author of the meta-analysis to obtain it. All retrieved pdfs were then inspected for final inclusion or exclusion.

To be included in the final sample, meta-analyses needed to meet the following criteria:

i) Tested a specified empirical question, not a methodological one. Meta-studies that, for example, report statistics on methodological characteristics of previous studies were excluded.

ii) Sought to answer such question based on results of primary studies that had pursued the same or a very similar question. This criterion was necessary because, if primary studies had been conceived for completely different hypotheses from that of the meta-analysis, then there is no reason to suppose that the effects extracted from them would be biased with respect to the meta-analysis' hypothesis. To this end, we excluded meta-studies that used data from studies designed for a completely different research question (for example, a study that pools the weight of a farm animal breed by taking values of weight from any previous study done on that breed). We also excluded meta-analyses whose primary objective was to test a meta-question using meta-regression (for example, a meta-analysis that assesses the sensitivity of various species to habitat degradation by comparing ecological surveys that were conducted in different habitats for purposes other than measuring the effects of habitat degradation).

iii) Identified primary studies via a dedicated literature search and selection. Meta-analyses that reanalyzed data sets compiled by previous meta-analyses were excluded.

iv) Produced a formal meta-analysis, i.e., a weighted summary of individual outcomes of primary studies. Studies that had only produced a systematic review (e.g., presented tabulated results but had not pooled results in a summary estimate) and studies that adopted semiquantitative methods, such as vote counting, were excluded. We excluded studies that, instead of summarizing the outcomes of different experiments, had recompiled a single dataset from primary sources, and thus do not report primary study-level data and statistics. We also excluded genome-wide association studies (GWAS) and meta-analyses of neuroimaging data (e.g., fMRI) because their methods and outcomes are too different to be effectively compared with those of standard meta-analyses.

v) The meta-analysis included at least five independent primary studies. When studies had conducted multiple meta-analyses or had partitioned the meta-analysis into different subgroups, with no attempt to an overall summary, then we selected the submeta-analysis that had the largest number of independent primary studies.

Whenever in doubt about the inclusion or exclusion status of a paper or whenever the pdf and supporting information did not include the meta-analysis's primary data, attempts were made to contact the author of the meta-analysis by email, to ask for clarifications and/or data. A first attempt was made to contact all authors, using an email with standard text but that was personalized by specifying the name of the author and the title of the meta-analysis of interest. All responses were then acknowledged and dealt with by direct email exchanges with D.F. Authors who did not reply the first time were sent a reminder 1 wk later, followed by another reminder 4 wk later. When no response was received, the study was excluded.

Fig. S1 reports a summary flow diagram of the selection process as it occurred in each discipline. All of the potentially relevant titles were inspected for 17 of the 22 disciplines. In the remaining 5 disciplines, a random sample of titles was inspected instead.

The initial selection of potentially relevant titles was conducted by D.F. Then pdfs were downloaded by an assistant (Annaelle Winand) who operated a first screening, excluding all titles that were obviously not systematic reviews or meta-analyses. All decisions about inclusion and exclusion of the remaining pdfs were made by D.F.

Data Extraction. For each primary study in each included meta-analysis we recorded the study's identification number, its reported effect size, and the measure of precision provided (i.e., confidence interval, SE, or N). In most cases, primary data were available directly in numerical form, either within the meta-analysis's forest plot or in tables. When the meta-analysis presented data exclusively in graphic form (i.e., forest plot with no numbers), primary data were extracted using a plot-digitizing software. The nature of the metric used by the meta-analysis was also recorded.

When the meta-analysis did not provide primary data in any form (and/or when the inclusion/exclusion status of the meta-analysis was in doubt), multiple attempts were made to contact the authors of the meta-analysis.

All primary data were extracted by a team of research assistants (i.e., Sophy Ouch, Sonia Jubinville, Mélanie Vanel, Fréderic Normandin, Annaelle Winand, Sébastien Richer, Felix Langevin, Hugo Vaillancourt, Asia Krupa, Julie Wong, Gabe Lewis, and Aditya Kumar) and subsequently inspected, corrected, and completed by D.F.

Study and Author Characteristics.

Sources of bibliometric data. For each primary datum reported in each meta-analysis we searched and retrieved available bibliometric information, using multiple techniques and databases. Each reference was searched in the Web of Science Core Collection (WOSCC) using, whenever possible, links that are provided automatically by the WOSCC to all cited references of a text (in our case, of the meta-analyses). If the primary study was present in the WOSCC database, its entire bibliometric record was downloaded, as well as the bibliometric profile of all its authors (described below). When the study was not available in the

WOSCC database, its entire bibliographic reference, as reported in the meta-analysis's publication, was recorded instead. Studies that were cited in meta-analyses as being unpublished were recorded as such. All meta-analyses had been obtained from the WOSCC, and therefore all had a corresponding bibliometric record that was also retrieved.

To ensure maximum completeness of information, records that did not appear to be accessible via direct link from the meta-analysis were searched again by hand in the WOSCC, using looser criteria (e.g., using key words in the title and/or authors); when such search was unsuccessful, author names and key words in the study's title were searched in the entire Web of Science (which includes multiple alternative databases: BIOSIS Citation Index, CABI: CAB Abstracts and Global Health, Current Contents Connect, Data Citation Index, Derwent Innovations Index, Inspec, KCI-Korean Journal Database, MEDLINE, Russian Science Citation Index, SciELO Citation Index, Zoological Record) and, if still unsuccessful, in Scopus. The corresponding bibliometric records, when available, were recorded.

Moreover, for all records for which bibliometric data were not available, we searched the name of the first author to identify, in the WOSCC, an alternative paper from which author bibliometric information could be extracted (these parameters are described below).

All primary study characteristics were recorded and searched multiple times by different research assistants, and results were inspected, corrected, and completed by D.F. Further details of each parameter collected and how variables tested in the study were derived are provided in *Supporting Information*.

For studies for which no identifier was available, as much information as possible was extracted by the bibliographic reference available in the meta-analysis. Information available in these bibliographic references usually included year of publication, number of authors, and name and type of source (i.e., book, thesis, unpublished data, conference proceeding, journal, personal communication, etc.).

Individual authors characteristics. The complete bibliometric profile of all authors of all studies for which a WOSCC record was available (including authors for which an alternative record had been retrieved, described above) was obtained using a disambiguation algorithm (44). This algorithm clusters papers recorded in the database around individual author names, using decision rules that weight multiple items of information available in the database records (e.g., first name, email, affiliation, etc.). Depending on the level of information available, the precision of this algorithm varies between 94.4% and 100% (16). From each corpus associated with each author in our sample, we calculated the following bibliometric parameters:

First publication year: Year of the first publication of each author (for journals covered by WOS). This information was used to estimate the career level of the author (parameter below).

Last publication year: Year of the last publication of each author (for journals covered by WOS).

Number of papers: Total number of publications (article, review, and letter) counted up to the year 2014.

Total citations: Total number of citations (excluding author self-citations) for all of the publications, considering citations received up to the year 2014.

Average citations: Mean citation score of the publications calculated as the ratio of total citations to total number of publications.

Average normalized citations: Mean field-normalized citation score of the publications. This is calculated by dividing each paper's citations by the mean number of citations received from papers in the same WOS Subject Categories and year and then averaging these normalized scores across all papers published by the author.

Average journal impact: Mean field-normalized citation score of all journals in which authors have published. This measure would be conceptually similar to taking the average Journal Impact Factor of an author but, unlike the latter, it is not restricted to a 2-y time window and is normalized by field.

Proportion of top 10 journals: Proportion of publications of an author that belong to the top 10% most cited papers of their WOS Subject Categories.

Country of author: Country was attributed based on the linkages between authors and affiliations recorded in all their papers available in the WOS. Because not all records have affiliation information, and because authors might report different affiliations throughout their careers, country was attributed based on a majority rule, taking into account all of the countries associated with the author in all of the author's publications. The country that we indicate for an author, in other words, corresponds to the best estimate of the place of most frequent activity of the author.

Full first name: This could be obtained by having at least one paper in the author's corpus that reported his/her complete name.

Number of retracted publications coauthored by the author.

Based on the sources of information above, we derived the following independent variables:

Gray literature vs. journal article: Any record that could be unambiguously attributed to sources other than a peer-reviewed journal article was classified as "gray literature." This category includes all personal communications, unpublished material, working papers, conference presentations, graduation thesis (any degree), book sources, reports, and patents. Whenever in doubt, the record was classified as journal article, to ensure a maximally conservative analysis.

Year of publication within meta-analysis: This corresponds to the publication year, rescaled to the oldest year within each meta-analysis. For gray literature and other records lacking bibliometric information, year was derived from manual search or indirect sources (e.g., from the bibliographic reference provided in the meta-analysis).

Citations received by each study.

US study vs. not: Any study or author that could be unambiguously attributed to the United States as opposed to any other country. Country attribution followed a priority rule: It was based on corresponding address (or any other reliable geographic information pertaining to a study) whenever this was available, and secondarily it was based on the country ascribed to the first author and, when this was also unavailable, to the country of last author.

Industry collaboration vs. other: Any study in which one or more authors indicated, as one of his/her affiliations, a private sector organization. This attribution is based on computerized routines that identify the private-organization status of an affiliation based on indicative acronyms or details in the reported addresses (e.g., "inc.," "ltd.," etc.) (45). Because not all industry partnerships may be formally indicated (or identifiable) in the addresses reported in WOS records, this variable yields a conservative estimate. Nonetheless, it is a more reliable indicator than sponsorship status indicated in acknowledgements, the proxy normally used to assess industry influence, because sponsorship

information in acknowledgment sections is available only since 2009 and may not reliably reflect industry influence, given that sponsors are usually required to play no role in study design and execution.

Publication policy of country of author: This was based on country of study, attributed following the same priority rule used for the US-study variable described above. In secondary analyses, we also tested this parameter, distinguishing countries attributed as first and last author. Policy classification was based on categories proposed by previous independent literature (16).

Author publication rate: Total number of papers divided by total number of years of publication activity (author's last publication year minus author's first publication year plus one).

Author total number of papers, total citations, average citations, average normalized citations, average journal impact, and proportion of top 10 journals, as described above.

Team size: Number of authors, as described above. When bibliometric data were missing, number of authors was extracted from the author list, when available, or excluded in all other cases.

Country-to-author ratio: Number of countries included in corresponding addresses list, divided by total number of authors.

Average distance between author addresses, expressed in hundreds of kilometers: Geographic distance was calculated following the methodology of ref. 46, based on a geocoding of affiliations covered in the Web of Science (45).

Career stage of author: Number of years occurring between an author's first publication and the year of publication of the paper included in our sample.

Female author vs. male author: Gender was attributed based on a combination of name and, whenever available, main country of author's activity. The majority of names were classified using an online service (genderapi.com), and unclassified records were later completed, to any extent possible, by hand, using nation-specific lists of baby names and similar sources. Names that could not be attributed reliably to one gender were classified as "unknown".

Retracted author vs. not: Dummy variable identifying authors that had coauthored at least one retracted paper. This estimation was based on the author bibliometric data above. Authors for which no bibliometric data were available were ascribed no retractions, making this a conservative estimate.

Data Preparation and Analyses.

Data standardization. Data from meta-analyses that used interconvertible metrics (i.e., Cohen's *d*, Hedges' *g*, correlation coefficient, Fisher's *z*, odds ratios and any metric whose description corresponded to one of these) were all transformed to log-odds ratios—these data were used in all main analyses. The remaining metrics were left untransformed. Whenever possible, we recorded as primary study's unit of precision (for subsequent weighting) the study's SE, which could in most cases be retrieved directly from the publication or recalculated from other data available in the meta-analysis (e.g., confidence intervals or sample size in the case of correlation coefficients). In a few meta-analyses that used unorthodox metrics and for which only sample size was available, precision was calculated as the inverse of the sample size. Primary studies whose outcomes appeared more than once within a meta-analysis were included only once, to ensure that each primary dataset consisted of independent studies.

Each dataset was subsequently standardized following previously established protocols (14). Specifically, from each meta-analytical primary data set obtained above we recalculated a random-effects weighted summary and subtracted this value from each primary effect size within that meta-analysis, essentially centering all primary outcomes by meta-analysis.

Coining by expected direction of primary outcomes' biases. Depending on the specific phenomenon studied and the metric used, the effects tested by studies in a meta-analysis might be expected to be positive or negative. Biases that lead to overestimation of effects would be expressed in the positive direction in the former case and in the negative direction in the latter. In previous studies we classified the direction of expected effects by hand (what we called the "expectation factor"), but the method was deemed to be inapplicable to most disciplines (14). However, it is reasonable to assume that in meta-analyses in which the pooled summary values are negative, the tested effect is negative and biases are therefore expressed in that direction. Therefore, 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. All analyses presented in figures refer to meta-analyses coined with a threshold of zero. Secondary results obtained with uncoined data, and data with a more conservative coining threshold (i.e., 1 SD of primary effect sizes, described below) are reported in Datasets S1–S5.

Analyses. All analyses reported in the main text were obtained by running meta-analyses at multiple levels. Each individual bias or risk factor was first tested for its effects within each meta-analysis, using simple meta-regression (the full set of meta-regression estimates and their SEs are reported in Dataset S1). The meta-regression slopes thus obtained were then summarized by a second-order meta-analysis, weighting by the inverse square of their respective SEs. This second-level meta-analysis assumed random effects, i.e., allowed each effect to vary at random across meta-analyses.

The robustness of our results was assessed by repeating all analyses on uncoined data (described above), data coined with a more conservative threshold (i.e., one negative SD of the log-odds ratio values in our sample, value equal to −1.11), meta-regressions not adjusted for study size, and standardized meta-regression estimates (i.e., each first-level meta-regression estimate was divided by the SD of its corresponding independent variable).

To assess the relative independence of biases measured by second-order meta-analysis, we also analyzed data using weighted multilevel regression, a method that yields similar but more conservative estimates (32). A complete discussion of this method and all results obtained with it can be found in *Supporting Information*.

Study characteristics. We downloaded the entire database record of all primary studies (and meta-analyses) for which a database identifier was available. From these records, using computational methods, the following study characteristics were extracted: (*i*) year of publication, (*ii*) total citations received at time of sampling (i.e., December 2013), (*iii*) number of authors, and (*iv*) corresponding address.

SI Limitations of Results

First, we assumed that the expected direction of an effect corresponds to the direction of the summary effect of the meta-analysis, because this is the best guess given the data. Following this assumption, all primary effect sizes within meta-analyses whose summary estimate was below zero had their signs inversed (a process known as coining). Our results were robust to relaxing this assumption by allowing meta-analyses not to be coined and were actually strengthened (i.e., yielded additional statistically significant results) if more conservative coining assumptions were made (Dataset S2). Nonetheless, it is likely that our analysis included

occasional errors, i.e., cases of meta-analyses in which the expected direction of effects was negative. Hand coding the direction of expected effects might have avoided some errors, but the procedure would be unfeasible and/or unavoidably subjective in many disciplines (14). Second, we tested all bias patterns and risk factors, assuming simple linear relationships and adjusting at most for one covariate (i.e., study precision). Multivariable analyses, however, support our general conclusions by showing that, once all predictors are included in the same model, the relative strengths of biases and risk factors are similar (*Supporting Information*). Third, our study relies on data reported by published meta-analyses, which could have errors or biases of their own. Empirical surveys suggest that such errors are common (47). Nondifferential measurement errors within meta-analysis are associated with deflated associations, so the impact of bias may be even larger in some cases. Fourth, we considered only meta-analyses with at least five studies, whereas it is common for many scientific topics to have only one or a few studies available (48). Fifth, the accuracy and internal consistency of data collected were very high for most variables (Table S6), but occasional measurement errors or missing data may have introduced noise that, if present, would lead to an underestimation of measured effects. Sixth, we assumed random effects in all second-order meta-analyses, an assumption that generalizes our findings at the cost of reducing statistical power, particularly in the presence of heterogeneity.

SI Multilevel Meta-Regression Analysis

Preliminary Technical Note. In a previous study with similar design but more limited scope (14) primary studies extracted from meta-analyses in Web of Science Subject Categories of "Psychiatry" and "Genetics & Heredity" were analyzed using a multilevel weighted regression approach. All independent variables (i.e., study characteristics) were standardized by meta-analysis and all primary outcomes were centered around their respective meta-analytical random-effects summary, to be analyzed as a nested population of individual data points, using inverse-variance weighted regression. The main analysis ran such a regression on a measure of how extreme a reported result is, which was called "deviation score," calculated as the absolute value of primary outcomes, square-root transformed twice to approximate normality. The model included a random intercept to account for differences in overall dispersion of data points between meta-analyses. The second main analysis assessed the effect of a binary variable encoding the expected direction of outcome in each meta-analysis (i.e., whether the desired effect was positive or negative, a measure we called expectation factor) on the nonabsolute values of standardized primary outcomes.

It was later shown that a reanalysis of the data using a simpler meta-meta–analytical approach (i.e., the method used in the present study) yielded results substantially similar in content and actually easier to interpret (32). The meta-meta–analytical approach has the advantage of yielding direct estimates of effects of bias within and across meta-analyses and also provides direct estimations of between–meta-analysis variance in effects. However, this method has the disadvantage of performing regressions on what are typically small numbers of studies per meta-analysis. This method therefore hampers multivariable regression analysis, because the number of variables that can be tested in any individual analysis is limited by the size of the smallest meta-analyses included. The average size of meta-analysis in the literature is rather small and our sample, which aimed at being representative of the population of meta-analyses at large, included any meta-analysis with $k \geq 5$.

The multilevel weighted regression approach avoids the limitation imposed by the small number of studies in the included meta-analyses, but requires standardization and transformations that make results less straightforward to interpret. Furthermore, multilevel regression accounts for heterogeneity at the lowest level of analysis (i.e., the level of primary studies), using a multiplicative factor (49) and not an additive factor (50). This difference makes results of multilevel models rather sensitive to distributional properties of data (normal distribution of errors is assumed at all levels of analysis) as well as of weights [large studies have more weight in a multivariable model than they would in a random-effects meta-analysis with additive random-effects correction (49)].

To summarize, the meta-meta–analytical approach presented in the main text estimates more closely and directly the amount of bias that may be encountered in meta-analysis. To assess the relative independence of multiple factors, however, multilevel weighted regression (henceforth, referred to simply as "regression") represents a more powerful approach. In *Supporting Information* we therefore follow this latter approach to test for the relative independence of biases and bias risk factors measured in the main text and run further secondary analyses.

Multilevel Analyses. Before proceeding with the analyses, we assessed the validity of the regression method by attempting to replicate previous results on the new sample (14). To this end, we selected from our new sample all meta-analyses that the Web of Science Subject Category system (i.e., not the Essential Science Indicators classification but the Subject Category classification) had classified under Psychiatry or Genetics & Heredity. To match the characteristics of the original sample, year of publication of meta-analysis was limited to the years 2009–2012 and size of meta-analysis to between 10 and 20 independent primary studies. We thus obtained a subsample of 42 behavioral and 29 nonbehavioral meta-analyses (852 studies in total), which is comparable in all characteristics with that tested previously.

To replicate the previous analysis, we analyzed the effects on deviation score (description above) of small-study effects, US effect, and year in meta-analysis. In our previous study we estimated confidence intervals by using a Markov chain Monte Carlo sampling algorithm, which, however, has been discontinued from the R lmer package that we used for these analyses (49). Therefore, confidence intervals are here derived by a simple Wald method. Because errors appear to be symmetrically heavy tailed in these data, confidence intervals estimated in this analysis are bound to be very conservative (i.e., statistical significance is underestimated).

Overall, results of this analysis are substantially similar to those previously reported (49), suggesting that relatively larger-effect sizes are observed by US studies among behavioral meta-analyses and not in the nonbehavioral ones (Table 1, leftmost columns). However, the values of this effect are shifted in the negative direction and confidence intervals are very large. Inspection of the data revealed the presence of a few meta-analyses with extremely low reported variances, which gives a few data points a vastly inflated weight in the analysis. The largest weight in the sample is 6.1×10^7, which represents over 63% of all weights in the sample combined—clearly an unacceptable imbalance.

The distribution of weights improved only slightly if SEs were rescaled and centered by meta-analysis, using the same formula as in previous analyses (14); i.e.,

$$\frac{t_{ij}}{\bar{t}_j}\hat{t},$$

where t_{ij} is SE of study i in meta-analysis j, $\bar{t}_j$ is the average SE in meta-analysis j, and $\hat{t}$ is the average of all SEs in the sample. This rescaling of SEs shifted results closer to those previously obtained (Table 1, center column), but still produced some extreme variance values—the largest weight being equal to 1.3×10^7, i.e., 26% of the total. To avoid the biasing effects of extreme

weights, we therefore rescaled by meta-analysis not the SEs, but the weights themselves, with the formula

$$\frac{w_{ij}}{\sum w_j} * \frac{k_j}{\sum k_j}$$

where w_{ij} is the inverse-variance–derived weight for study i in meta-analysis j (i.e., t_{ij}^{-2}, in the notation above), $\sum w_j$ is the sum of all weights in meta-analysis j, k_j is the size (number of primary studies) of meta-analysis j, and $\sum k_j$ is the sum of all primary studies in the sample. The formula simply rescales weights within meta-analysis and the total weight of all studies in each meta-analysis is proportional to the number of studies that this meta-analysis has. With this transformation, the largest weight in the sample is 0.003, i.e., 0.3% of the total, and results of the regression analysis (Table S2, rightmost columns) replicate our previous results very closely in magnitude and direction as well as in confidence interval values. It is worth pointing out that a reanalysis of previous data (14) with these rescaled weights yielded analogous results to those originally reported.

When this latter analysis (i.e., with rescaled weights) was extended to the rest of the new sample (1,864 usable meta-analyses for a total of 24,622 nonnull data points), results corroborated those obtained with meta-meta–analysis and reported in the main text, showing a substantive small-study effect, a significant US effect across all disciplines, and no early-extremes effect (i.e., regression estimates with rescaled weights: study size = 0.202 [0.189, 0.215], US = 0.010 [0.004, 0.015], year in meta-analysis = 0.000 [−0.002, 0.003]). We did not code meta-analyses for an expectation factor and cannot therefore replicate the second part of our original analyses. Nonetheless, if applied to coined standardized primary effect sizes (i.e., the same outcomes used in the main text), analysis on the whole sample yields a substantive small-study effect (b = 0.589 [0.508, 0.670]), a marginally nonsignificant US effect (b = 0.016 [−0.017, 0.049]), and little to no decline effect (b = −0.004 [−0.021, 0.013]).

Therefore, in line with previous results (14, 32), we conclude that a multilevel regression with random intercept yields substantially similar but more conservative results than the meta-meta–analytical approach adopted in the main text, once weights are rescaled by meta-analysis, as illustrated above.

We use this latter approach to assess the relative independence of the biases and risk factors described in the main text. Each analysis was repeated using three different standardizations of primary study outcomes: "effect sizes" (i.e., original outcomes, standardized by meta-analysis), "coined effect sizes" (i.e., the same as effect sizes, but with their sign inverted when the overall weighted summary of their respective meta-analysis was negative; see *Methods* for further details), and deviation score (i.e., double square-root–transformed absolute value of reported primary outcomes, which is a measure of how extreme results are in either direction). All continuous parameters tested in these models were rescaled by meta-analysis with z transformation.

Multivariable test of citation bias. The number of citations received by a study (which for this analysis is rescaled by meta-analysis via z transformation) was significantly predicted with the magnitude of its effects when tested alone (e.g., on coined effect sizes: b: 0.016 [0.006, 0.027]) as well as when adjusted by study size and by year of study (standardized by meta-analysis) (Table S3), suggesting that citation bias is independent of small-study effects as well as decline effects. Deviation scores, however, were negatively associated with citations (Table S3, rightmost column). This strongly suggests that studies do not get more citations in proportion to how extreme they are in general, but in proportion to how extreme they are in the direction predicted by the hypothesis—qualifying this effect as a true "bias." This interpretation is further supported by the fact that coined effect sizes (which are expectation-corrected effect sizes, Table S3, center) were more strongly associated with citations than normal effect sizes, which had a positive but nonsignificant effect.

Interestingly, in addition to the citation bias determined by a study's reported effect sizes, we also detect a negative effect on citations of a study's SE, which suggests that, once adjusted for effect size magnitude, larger, more precise studies receive more citations. We also consistently observed a negative effect of year of study, which is expected given that older studies had more time to cumulate citations (Table S3).

Relative independence of biases. We tested for all other main biases reported in the main text. Analyses confirmed our main results, showing that small-study effects are substantially larger than any other form of bias and that gray literature bias is second by importance. All other biases have smaller magnitudes, but are generally measured in the direction predicted and shown to be independent of each other, with the exception of sponsorship bias, the direction of which was sensitive to analytical choices (Table S4 and Dataset S2).

Finally, we tested all of the main risk factors, adjusting for study's SE. Again, analyses confirm results obtained with univariate meta-meta–analysis (Table S5 and Fig. 1), suggesting a prominent effect of study size, a negative association with team size, a positive association with country-to-author ratio, and largely null or negative effect for most other parameters except that of having a retracted first author. Note how effects on deviation score are in most cases null and effects on expectation-corrected effect sizes are generally stronger than in uncorrected effect sizes (Table S5, center column vs. left column, respectively), supporting our assumption that, to any extent that they have significant effects, these factors increase the risk of directional biases—i.e., they are all biases in the literal sense of the term.

Relative strength of biases. Our main analyses suggested that biases vary widely in their relative importance. This claim was based on the assessment of level-2 random-effects summaries of a meta-regression coefficient estimated using various models and measured directly (i.e., b values) or in a standardized fashion (i.e., beta). Our regression approach further showed that similar relative strengths are obtained in a multivariable model, in which the different biases are included as z-transformed predictors (i.e., the independent variables are rescaled by their SD). To further assess the relative strength of each predictor, we measured their respective effect sizes, expressed as amount of variance explained (i.e., the coefficient of determination or R^2). Obtaining such measures in mixed-effects models is not straightforward, but a method to calculate a pseudo-R^2 was recently developed (51). For each variable tested in this study, we measured this pseudo-R^2 in the multilevel regression model as well as a standard R^2 obtained in a simple linear and weighted regression (i.e., with the random intercept omitted).

As illustrated in Table S6, whereas the exact estimation of variance explained varies greatly depending on the model adopted, results consistently show small-study effects to be around 20 times larger than any other bias effect measured. The second largest is gray literature bias and the third is citation bias, if citations are tested as predictors of effect size. If, however, citations are taken to be the independent variable and effect sizes to be their predictors (as analyzed in Table S3, but not in Fig. 1), then citation bias is second in importance.

Secondary tests of early extremes, Proteus phenomenon and decline effect. Our main analyses assessed for the possible reporting of more extreme effects among earlier studies within a meta-analysis and found little support for it. A subtler variant of this hypothesis is represented by what has been coined the "Proteus phenomenon," in which the first study published about a specified effect reports an extreme effect in the direction predicted and this event opens up a window of opportunity to publish in a very short time a study that reports a diametrically opposite effect, i.e., null or negative (19). This phenomenon is likely to be observed in fields in which results can be produced and published

very rapidly, even within the same year. Nonetheless, we conducted an exploratory test for the possible presence of the Proteus phenomenon by contrasting effects reported during the first year of the meta-analysis vs. those reported in the second and third years, adjusting for year of inclusion in meta-analysis (i.e., adjusting for a decline effect). Across the sample, results reported during the first year were significantly larger than their expected (decline-effect adjusted) value (b [95% CI] = 0.077 [0.022, 0.132]) whereas results of the second and third years also tended to be larger than expected, albeit not at formally statistically significant levels (0.03 [−0.021, 0.081]). Adjusting for study size in addition to year of inclusion yielded a similar effect for the first year (0.062 [0.007, 0.116]), but a smaller (null) effect for second and third years (0.008 [−0.043, 0.059]). Throughout these analyses, the variable representing year of meta-analysis continued to have a null or positive effect, against the hypothesis of a generalized decline effect.

These secondary results do not support the widespread (cross-disciplinary) presence of early-extreme effects and a Proteus phenomenon as measured here. It should be emphasized, however, that our tests, based on separating studies by year, are rather coarse grained and that the Proteus phenomenon as originally described should be expected to occur only in fields with particularly fast publication rates.

However, these results yield further insights into the decline effect, which our main findings refuted. These secondary analyses suggest that, whereas there is no universal "decline effect" in the literal sense, there is a strong "first-year" effect, in which earlier studies are simply more likely to overestimate the overall effect than all later ones.

Reliability and Consistency of Collected Data. In 47 meta-analyses in our sample, data were collected and processed twice independently for all variables measured. For each of these pairs of meta-analyses, we calculated the level of similarity for each variable used in our main analyses. Similarity was measured by Spearman's correlation for continuous variables and by proportion of matches for categorical variables. Table S5 reports the mean values obtained across meta-analyses for each of these variables and the average proportion of missing data per meta-analysis. As illustrated, the average correlation/matching between independently collected datasets was always above 98% for all variables, indicating that data collected were highly internally consistent. With the exception of gender attribution, which could not be done reliably on Southeast Asian names, 9% or less of the data were missing in any meta-analysis, and multiple variables had ∼0% missing data.

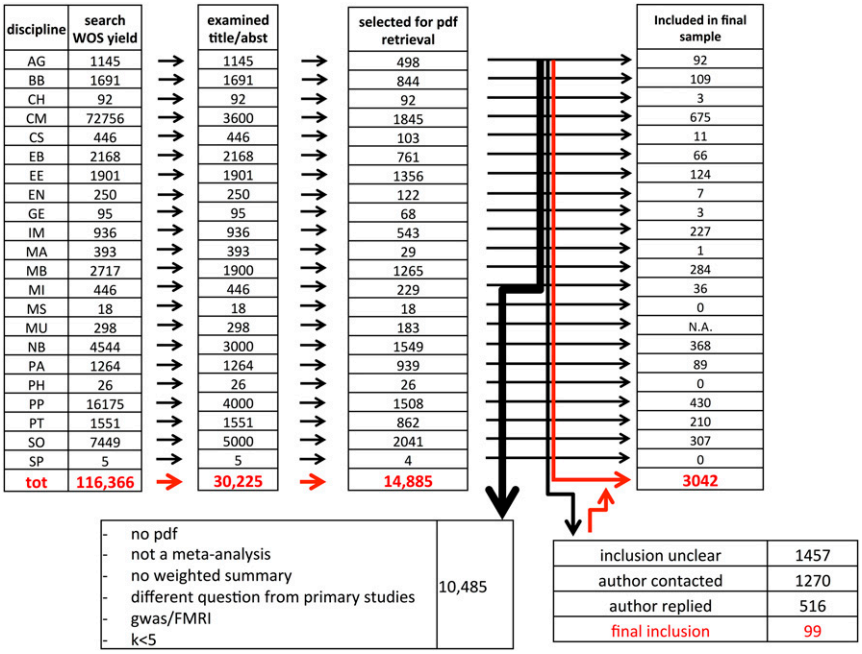

Fig. S1. Flow diagram illustrating the stages of selection of meta-analyses to include in the study. Electronic search in the Web of Science yielded an initial list of potentially relevant titles. From this list 30,225 titles were actually screened and, in successive phases of selection, a total of 14,885 titles were identified for potential inclusion. A total of 10,485 studies were eventually excluded for failing to meet one or more of the criteria, whereas 1,457 were deemed of unclear status. Of the latter, 187 were identified too late in the process to be included, whereas multiple attempts were made to contact the authors of the 1,270 previously identified studies. The authors of 516 of these studies responded, and communications with them led to discarding 417 studies, which either did not meet our inclusion criteria or had unretrievable data. Therefore, 99 meta-analyses could be included, bringing our final sample to a total of 3,042 distinct meta-analyses. Meta-analyses in the multidisciplinary category (MU) were reclassified by hand.

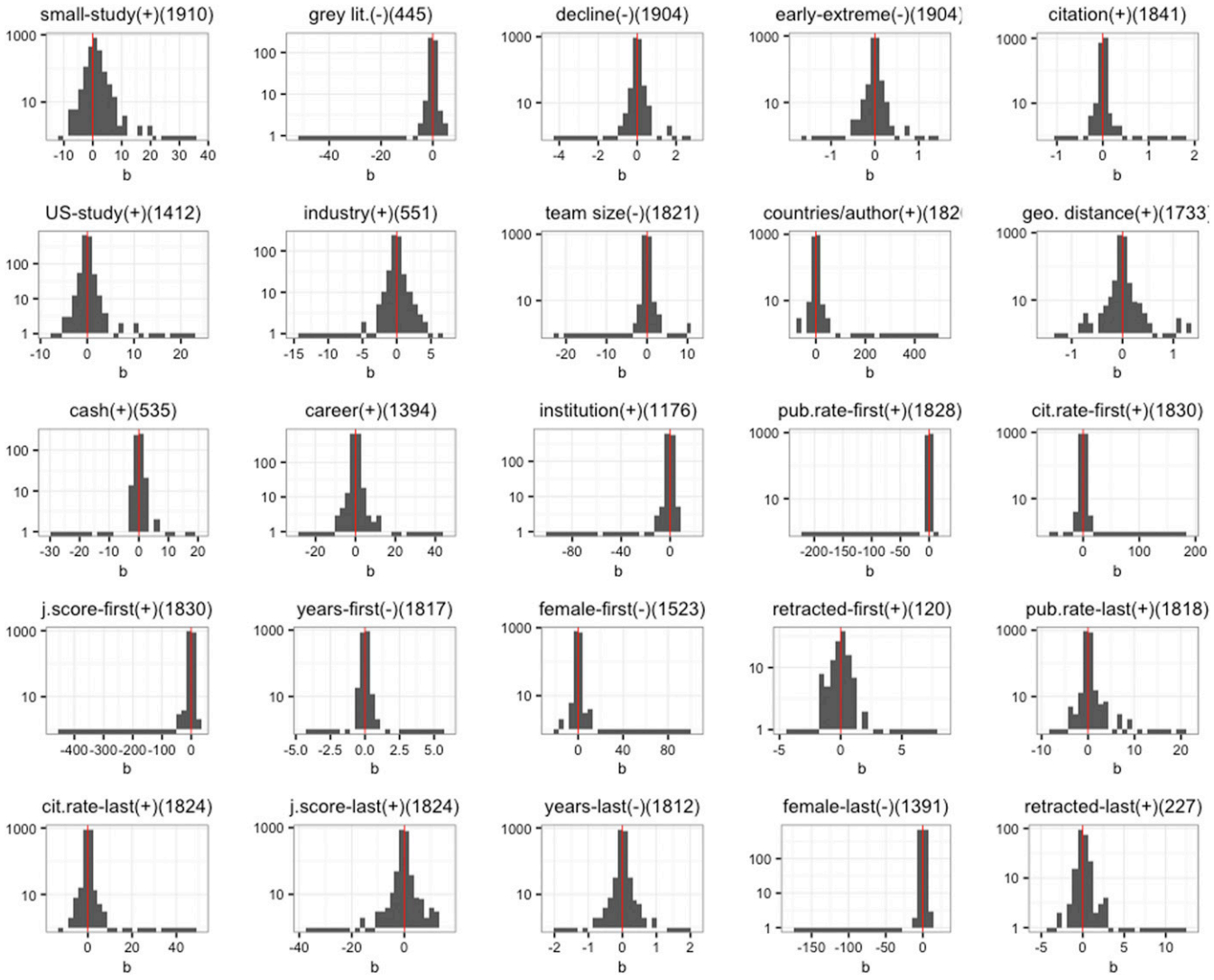

Fig. S2. Histograms of meta-regression coefficients measured for each bias pattern and risk factor tested in the study. Symbols in parentheses indicate whether the association between factor and effect size is predictive to be positive (+) or negative (–). Numbers in parentheses indicate the number of meta-analyses that could be used in the analysis (i.e., in which the corresponding independent variable had nonzero variance). Histograms are centered on zero, marked by a vertical red bar.

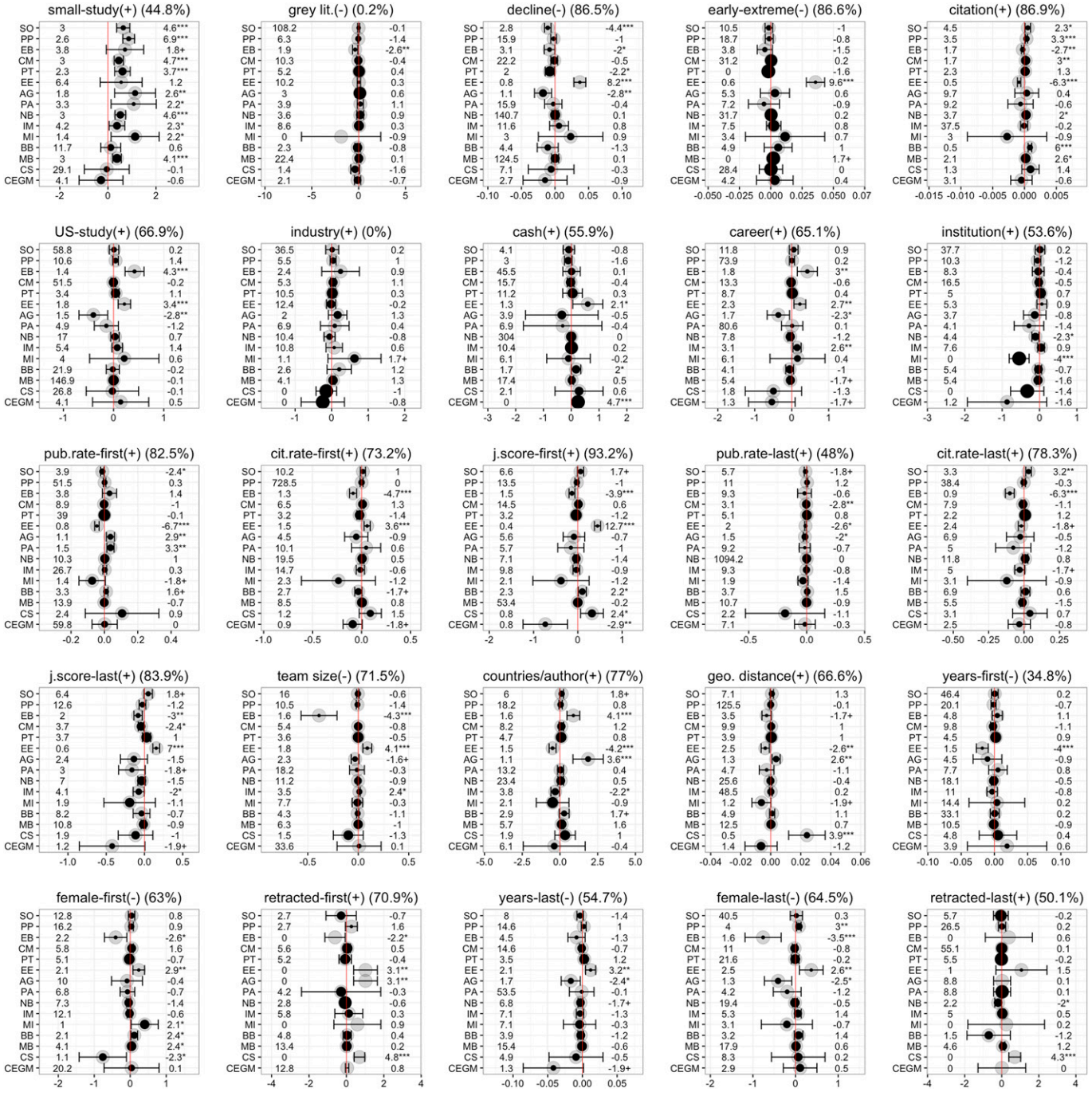

Fig. S3. Bias patterns and bias risk factors tested in our study, partitioned by discipline (see *Methods* for details). Each panel reports the second-order random-effects meta-analytical summaries of meta-regression estimates ($b \pm 95\%$ CI) measured across the sample of meta-analyses. The shaded portion of the area of each circle is proportional to the percentage of total variance that is explained by between–meta-analysis variance (i.e., heterogeneity, measured by I^2, all numerical results available as *Supporting Information*). Symbols in parentheses indicate whether the association between factor and effect size is predictive to be positive (+) or negative (–). Percentages above each panel report between-discipline heterogeneity. Meta-analyses in the "multidisciplinary" category were reclassified by hand in one of the other disciplines, and those from the physical sciences disciplines, being few in number, were combined in one category. Abbreviations are defined in *Methods* in the main text. CEGM, Chemistry + Engineering + Geosciences + Mathematics. To help visualize effects, numbers on the right of error bars display *t* scores and statistical significance levels (i.e., °$P < 0.1$, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). Numbers on the left of each error bar reflect the cross–meta-analytical consistency of effects, measured as the ratio of between–meta-analysis variance divided by summary effect size (i.e., τ^2/b; the smaller the ratio is, the higher the consistency).

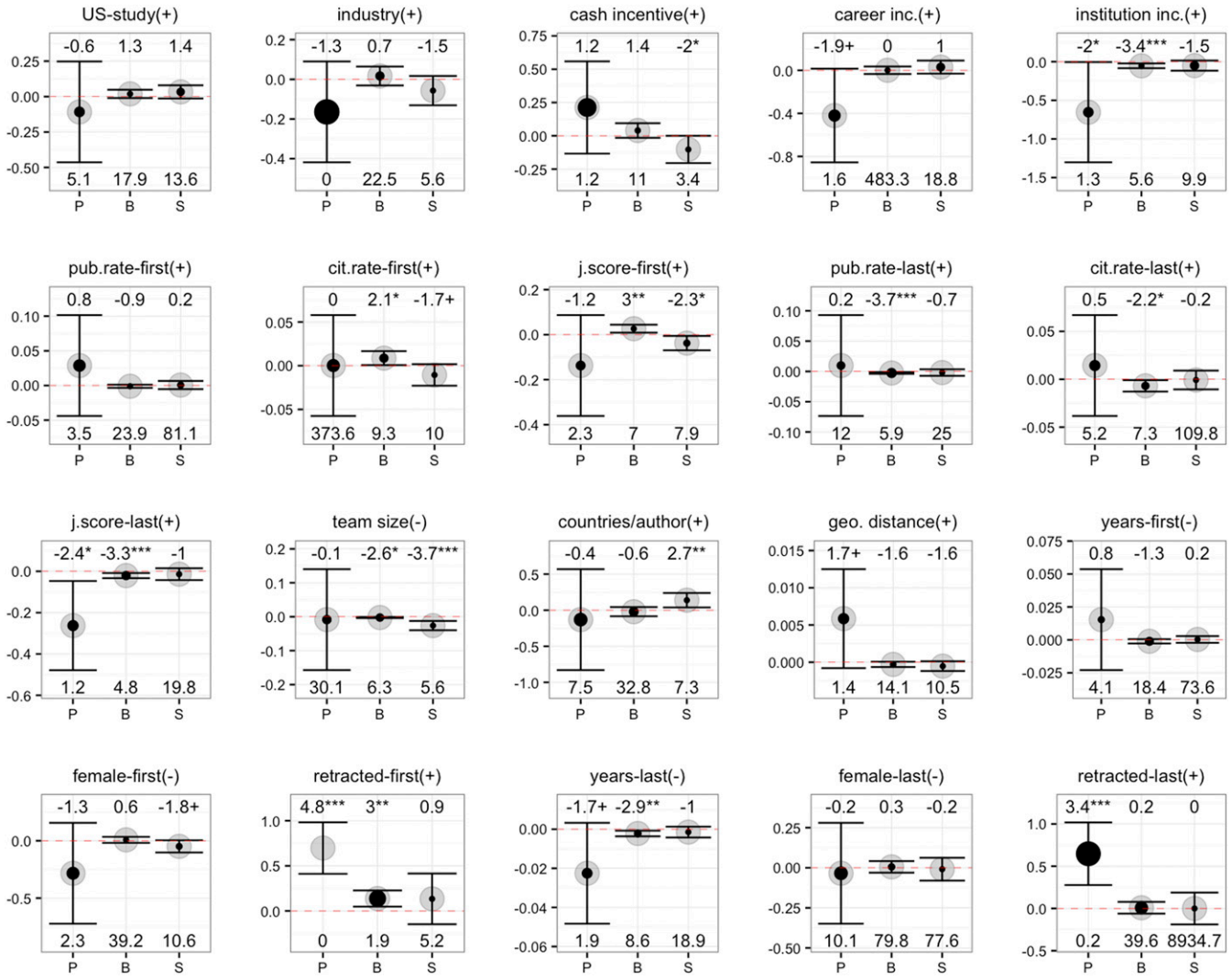

Fig. S4. Bias risk factors tested in our study, partitioned by disciplinary domain. Each panel reports the second-order random-effects meta-analytical summaries of meta-regression estimates ($b \pm 95\%$ CI) measured across the sample of meta-analyses. Symbols in parentheses indicate whether the association between factor and effect size is predictive to be positive (+) or negative (–). The transparent portion of the area of each circle is proportional to the percentage of total variance that is explained by between–meta-analysis variance (i.e., heterogeneity, measured by I^2, all numerical results available as *Supporting Information*). The sample was partitioned between meta-analyses from journals in the physical (P), biological (B), and social (S) sciences, as classified by Thompson Reuters' Essential Science Indicators database. To help visualize effects, numbers above error bars display t scores and statistical significance levels (i.e., $^{\circ}P < 0.1$, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$). Numbers below each error bar reflect the cross–meta-analytical consistency of effects, measured as the ratio of between–meta-analysis variance divided by summary effect size (i.e., τ^2/b; the smaller the ratio is, the higher the consistency). Using abbreviations described in *Methods*, discipline classification is the following: physical sciences (P), MA, PH, CH, GE, EN, CS ; social sciences (S), EB, PP, SO; and biological sciences (B), all other disciplines.

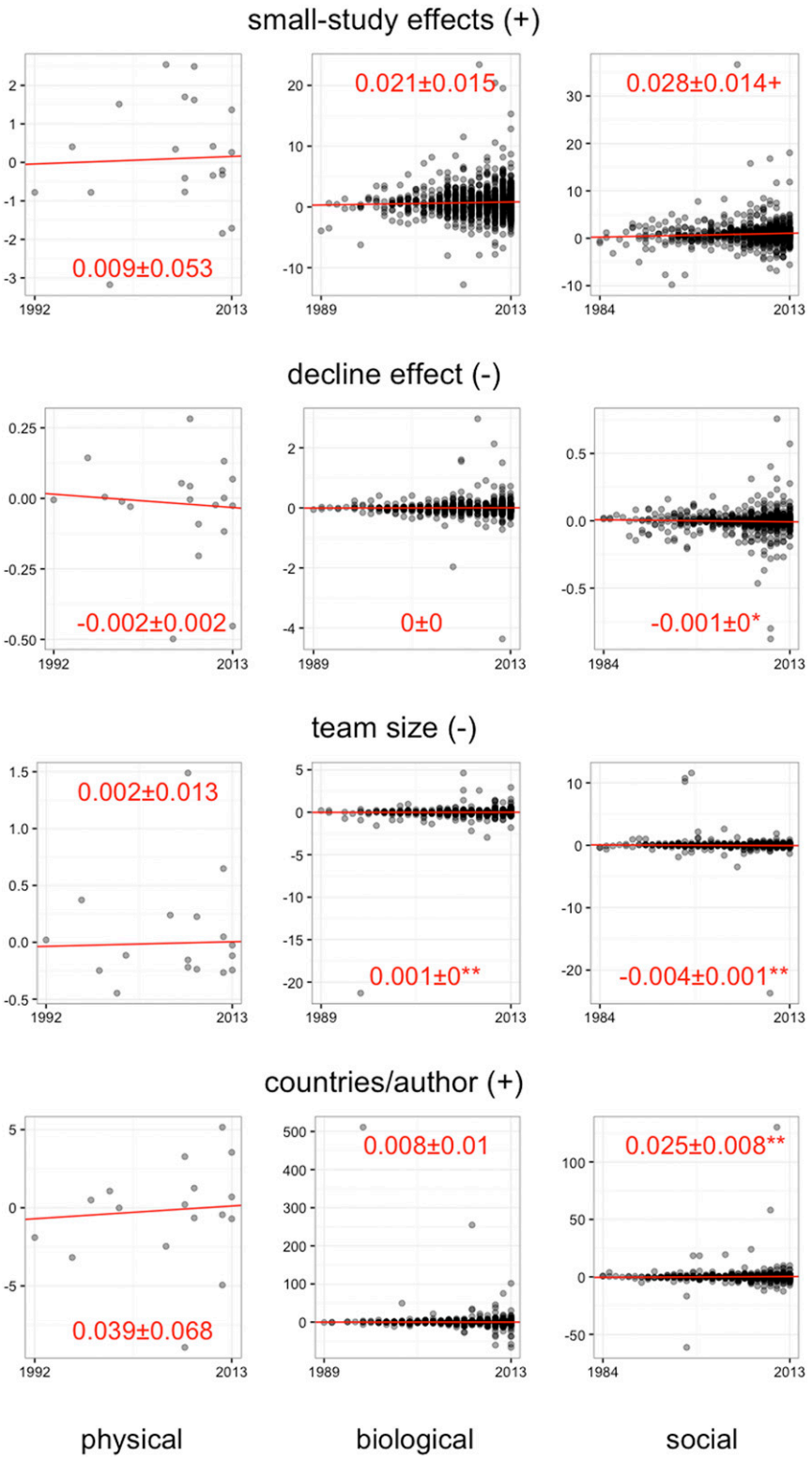

Fig. S5. Estimate of changes over time of three of the bias patterns tested (all numerical results are in Dataset S4). Circles in each panel are meta-regression estimates plotted by year of the corresponding meta-analysis. Regression lines and values in red were produced by a second-order mixed-effects meta-regression. Symbols in parentheses indicate whether the association between factor and effect size is predictive to be positive (+) or negative (–). See main text and Table 1 for further details. To help visualize trends, asterisks display conventional statistical significance levels (i.e., $+P < 0.1$, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$). Using abbreviations described in *Methods*, discipline classification is the following: physical sciences (P), MA, PH, CH, GE, EN, CS; social sciences (S), EB, PP, SO; and biological sciences (B), all other disciplines.

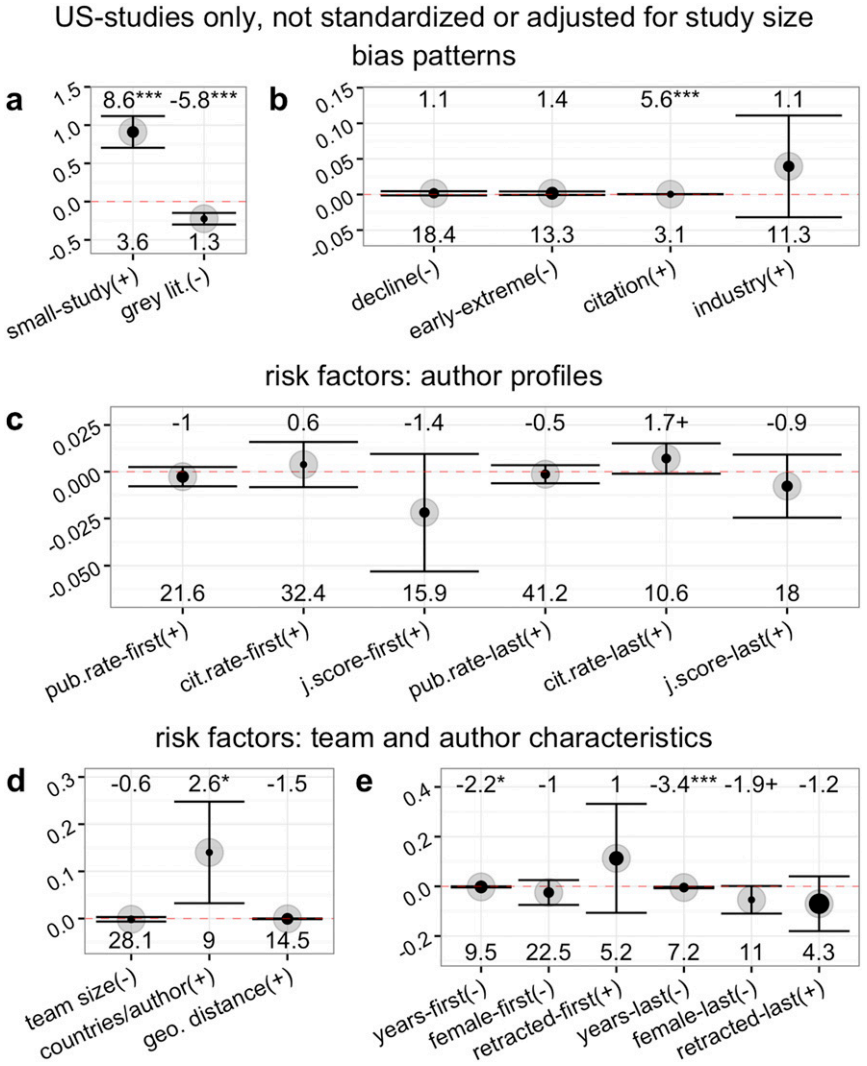

Fig. S6. Meta-meta–regression estimates of bias patterns and bias risk factors, not adjusted for study precision, limited to US studies. (*A*–*E*) The figure reports exactly the same analysis as reported in Fig. 1, with the exception of *B*, from which the test of the US effect, which does not apply to this sub-analysis, was removed. Details of each analysis and panel are equivalent to those reported in the main text. Each panel plots the second-order random-effects meta-analytical summaries of standardized meta-regression estimates (i.e., beta ± 95% CI), measured across the sample of meta-analyses. Symbols in parentheses indicate whether the association between factor and effect size is predictive to be positive (+) or negative (–). The gray area within each circle is proportional to the percentage of total variance explained by between–meta-analysis variance (i.e., heterogeneity, measured by I^2). To help visualize effect sizes and statistical significance, numbers above error bars display *t* scores (i.e., summary effect size divided by its corresponding SE, *b*/SE) and conventional significance levels (i.e., $+P < 0.1$, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$). Numbers below each error bar reflect the cross–meta-analytical consistency of effects, measured as the ratio of between–meta-analysis SD divided by summary effect size (i.e., τ/b; the smaller the ratio is, the higher the consistency). A few of the tested variables were omitted from *D* for brevity (full numerical results for all variables are in Dataset S2). See main text and Table 1 for details of each variable.

Table S1. Metrics used by the meta-analyses included in the study and their frequency in each of the 22 disciplines used for sampling

Metric	MA	GE	CS	CH	EN	MB	BB	MI	NB	PA	EE	IM	PT	CM	AG	PP	EB	SO	Total
Odds ratio	1	0	1	2	3	255	42	12	153	10	23	105	83	213	7	59	2	67	1,038
Risk ratio	0	0	0	0	0	14	21	15	42	10	29	39	59	173	30	26	1	40	499
Stand.mean.diff.	0	0	2	0	0	3	11	0	70	13	22	16	14	56	14	165	8	82	476
Correlation coefficient	0	1	6	0	1	0	3	0	16	23	12	2	1	10	2	95	39	42	253
Mean difference	0	0	0	0	0	3	17	0	27	7	4	11	18	80	27	7	1	16	218
Proportion	0	0	0	0	1	2	5	7	21	2	2	40	12	78	1	14	1	15	201
Hedges' *g*	0	1	2	0	0	0	2	1	17	9	2	0	3	7	2	42	0	24	112
Hazard ratio	0	0	0	0	0	3	2	0	2	0	0	4	3	31	1	2	0	3	51
Prop.diff.	0	1	0	0	1	0	1	0	5	2	2	2	5	14	2	4	3	3	45
Fisher's *Z*	0	0	0	0	0	1	1	0	2	2	8	0	0	1	0	8	0	1	24
Simple mean	0	0	0	0	0	1	0	1	2	4	3	1	2	3	0	0	0	0	17
Other	0	0	0	1	1	2	4	0	11	7	17	7	10	9	6	8	11	14	108
Total	1	3	11	3	7	284	109	36	368	89	124	227	210	675	92	430	66	307	3,042

Abbreviations of column headings are defined in *Methods* in the main text. Stand.mean.diff, standardized mean difference; Prop.diff, proportion difference.

Table S2. Measures of US effect using different weighting methods: Multilevel weighted regression estimates of the effects on deviation score of a study size (i.e., SE centered by meta-analysis), its chronological order of appearance within the meta-analysis (*z* scaled by meta-analysis), and the geographical origin of its corresponding author (United States vs. all other countries)

	Original variance		Rescaled variance		Rescaled weights	
Variable	GH	PP	GH	PP	GH	PP
(Intercept)	0.554 [0.506, 0.601]	0.692 [0.641, 0.743]	0.534 [0.486, 0.582]	0.658 [0.604, 0.711]	0.534 [0.484, 0.584]	0.601 [0.546, 0.656]
USA vs. rest	−0.067 [−0.104, −0.030]	−0.025 [−0.056, 0.007]	−0.042 [−0.087, 0.003]	0.024 [−0.013, 0.062]	−0.047 [−0.093, −0.001]	0.074 [0.037, 0.112]
Study size (SE centered and rescaled by meta-analysis)	0.219 [0.137, 0.301]	0.121 [0.042, 0.199]	0.293 [0.199, 0.387]	0.203 [0.110, 0.296]	0.302 [0.207, 0.398]	0.295 [0.199, 0.392]
Year in MA (*z* transformed by meta-analysis)	−0.015 [−0.03, 0.001]	−0.003 [−0.019, 0.013]	−0.003 [−0.020, 0.014]	−0.002 [−0.020, 0.016]	−0.005 [−0.022, 0.012]	0.006 [−0.012, 0.023]

Values are main effects and 95% CIs, obtained with different methods of weighting: i.e., using originally given primary study variances, variances rescaled to be centered by meta-analyses, and centered weights. Analyses are partitioned in two subsamples: GH meta-analyses ($k = 28$, $n = 351$) and PP meta-analyses ($k = 42$, $n = 499$).

Table S3. Citation bias: Multilevel weighted regression estimates of the effects on citations received by primary studies (rescaled by meta-analysis via *z* transformation) of study's reported effect size (standardized by meta-analysis as described in *Methods*), study size (i.e., SE centered by meta-analysis), and its chronological order of appearance within the meta-analysis (*z* scaled by meta-analysis)

Variable	Effect size	Coined effect size	Deviation score
(Intercept)	0.323 [0.294, 0.352]	0.328 [0.299, 0.358]	0.380 [0.332, 0.429]
Effect size*	0.005 [−0.005, 0.012]	0.021 [0.011, 0.031]	−0.087 [−0.145, −0.029]
Study size (SE)	−0.592 [−0.656, −0.528]	−0.604 [−0.668, −0.539]	−0.575 [−0.640, −0.510]
Year in MA	−0.347 [−0.359, −0.334]	−0.347 [−0.359, −0.334]	−0.347 [−0.359, −0.334]

Values are main effects and 95% confidence intervals. Each column represents the same model, but with different standardizations of primary study outcomes (i.e.noncoined and not trasforned effect size, coined effect size, double-square-root absolute value of effect size, see text for further details).
*Measured as indicated in each column heading.

Table S4. Bias effects, multivariable test: Multilevel weighted regression estimates of the effects on primary study's reported effect size of study size (i.e., SE centered by meta-analysis), gray vs. nongray literature status (binary variable coding whether the study was published in any outlet other than a peer-reviewed journal), decline effect (i.e., chronological order of appearance within the meta-analysis, by year *z* scaled by meta-analysis), US effect (dummy variable separating studies with corresponding or main author from the United States vs. any other country), and industry sponsorship effect (dummy variable separating industry-sponsored studies vs. all others)

Variable	Effect size	Coined effect sizes	Deviation score
(Intercept)	−0.094 [−0.127, −0.061]	−0.262 [−0.295, −0.229]	0.660 [0.652, 0.669]
Small-study effect	0.197 [0133, 0.264]	0.530 [0.464, 0.595]	0.183 [0.172, 0.194]
Gray literature bias	−0.092 [−0.143, −0.041]	−0.104 [−0.155, −0.053]	0.008 [−0.001,0.016]
Decline effect (or early extremes)	−0.012 [−0.025, 0.001]	−0.005 [−0.018, 0.008]	−0.002* [-0.005, 0.000]
US effect	0.025 [−0.004, 0.053]	0.013 [−0.016, 0.041]	0.003 [−0.001 to 0.008]
Industry	0.018 [−0.044 to 0.081]	−0.011 [−0.073, 0.052]	−0.018 [−0.029, −0.007]

Values are main effects and 95% confidence intervals. Each column represents the same model, but with different standardizations of primary study's reported effect size (see text for details).
*With deviation score, the decline effect corresponds to the early-extremes effect.

Table S5. Bias risk factors, multivariable test: Multilevel weighted regression estimates of the effects on primary study's reported effect size of study size (i.e., SE centered by meta-analysis), team size (number of authors), country-to-author ratio (based on addresses reported in study), cash incentives (binary variable identifying studies from countries where publication performance is rewarded with cash), institutional incentives (binary variable identifying studies from countries where institutions are rewarded based on publication performance of their employees), years occurring between the first publication of the study's first author and the year of publication of study, average number of papers published by first author per year, average normalized citations received by first author, average normalized journal score of first author, and binary variables reflecting whether the first author is female and whether she has ever coauthored a retracted paper

Variable	Effect size	Coined effect sizes	Deviation score
(Intercept)	−0.096 [−0.138, −0.055]	−0.275 [−0.317, −0.234]	0.655 [0.645, 0.664]
Study size	0.204 [0.119, 0.290]	0.586 [0.501, 0.671]	0.201 [0.187, 0.215]
Team size	−0.006 [−0.025, 0.014]	0.004 [−0.016, 0.023]	0.000 [−0.003, 0.003]
Countries/author	0.010 [−0.010, 0.030]	0.019 [−0.002, 0.039]	−0.001 [−0.004, 0.002]
Cash incentives	0.047 [−0.030, 0.123]	0.056 [−0.020, 0.133]	−0.017 [−0.030, −0.005]
Institutional incentives	−0.010 [−0.052, 0.033]	−0.018 [−0.060, 0.023]	−0.004 [−0.010, 0.001]
Years of activity, first author	−0.010 [−0.027, 0.007]	−0.004 [−0.021, 0.013]	−0.002 [−0.005, 0.000]
Publication rate, first author	−0.003 [−0.020, 0.014]	−0.006 [−0.023, 0.010]	0.00 [−0.003, 0.002]
Citation rate, first author	0.006 [−0.015, 0.027]	0.019 [−0.002, 0.040]	0.000 [−0.004, 0.003]
Journal score, first author	−0.002 [−0.023, 0.019]	−0.018 [−0.039, 0.004]	0.00 [−0.007, −0.001]
Female first author	0.029 [−0.008, 0.066]	−0.033 [−0.070, 0.004]	−0.001 [−0.006, 0.005]
Retracted first author	0.190 [0.009, 0.371]	0.255 [0.074, 0.435]	0.035 [0.008, 0.063]
Years of activity, last author	−0.004 [−0.022, 0.014]	−0.011 [−0.028, 0.007]	−0.001 [−0.004, 0.001]
Publication rate, last author	0.002 [−0.016, 0.019]	0.005 [−0.012, 0.022]	−0.003 [−0.006, −0.001]
Citation rate, last author	−0.019 [−0.037, −0.002]	0.000 [−0.017, 0.017]	0.001 [−0.002, 0.003]
Journal score, last author	0.008 [−0.009, 0.026]	−0.006 [−0.023, 0.012]	−0.004 [−0.007, −0.001]
Female last author	0.047 [0.006, 0.088]	0.033 [−0.008, 0.074]	0.005 [−0.002, 0.011]
Retracted last author	0.021 [−0.115, 0.157]	0.033 [−0.103, 0.168]	0.018 [−0.003, 0.039]

The lower part of the table reports results for the same model, but with data from last author instead. Further details for all variables are given in *Methods* and in the main text. Values are main effects and 95% CIs. Each column represents the same model, but with different standardizations of primary study's reported effect size (see main text for details).

Table S6. Relative magnitude of biases, measured by variance explained: Comparison of effect size values of different biases, measured by each bias's ability to explain the variance of primary study's effect sizes

Variable	Pseudo-R^2	Standard R^2	Standard R^2, unweighted
Small-study effect	0.27145	0.0055	0.03507
Gray literature bias	0.01197	0.00029	0.00045
Citation bias (effect size as independent variable)	0.01897	0.0003	0.00103
Citation bias (citations as independent variable)	0.00545	0.0003	0.00103
Decline effect	0.00432	0.00023	0.00053
Early extremes (same as decline effect, but on deviation score)	0.00199	0.00138	0.0014
Industry	0.00038	0.00004	0.00008
US effect	0.00036	0.00005	0.0000004

For each variable values of total variance explained via pseudo-R^2 calculated on the multilevel regression model (i.e., weighted mixed-effects model), standard R^2 values obtained from simple linear weighted regression, and standard R^2 of a linear and non-weighted regression are shown.

Table S7. Estimation of consistency and completeness of primary data: Average consistency of data collected, estimated as Spearman rank correlation (for continuous and ordinal variables) or proportion of matching values (for nominal categories)

Variable	Correlation/overlap	Missing	*N* of pairs
Effect size	1	0	47
SE/sample size	1	0	47
Citation	0.98	0.07	46
Publication year	0.98	0.01	47
Team size	0.97	0.09	46
Countries/author	0.98	0.09	46
Geographic distance	0.99	0.09	46
Years, first	0.98	0.07	46
Publication rate, first	0.98	0.06	46
Citation rate, first	0.99	0.06	46
j score, first	0.99	0.06	46
Years, last	0.98	0.09	46
Publication rate, last	0.99	0.09	46
Citation rate, last	0.98	0.09	46
j score, last	0.99	0.09	46
Gray literature	0.99	0	47
US study	1	0.06	47
Industry	0.99	0	47
Female, first	1	0.2	46
Retracted, first	1	0.06	46
Cash	1	0.06	47
Career	0.99	0.06	47
Institution	0.99	0.06	47
Female, last	1	0.24	46
Retracted, last	1	0.09	46

The values reported are weighted means of correlations obtained from pairs of meta-analyses for which all data were collected twice independently. *N* of pairs: number of pairs of meta-analyses with nonmissing and nonzero variance, on which similarity indexes could be estimated.

Other Supporting Information Files

Dataset S1 (TXT)
Dataset S2 (TXT)
Dataset S3 (TXT)
Dataset S4 (TXT)
Dataset S5 (TXT)