Estimating Publication Bias in Meta-Analyses: A Meta-Meta-Analysis Across Disciplines and Journal Tiers

Maya B. Mathur[1] and Tyler J. VanderWeele[2]

[1]Quantitative Sciences Unit, Stanford University, Palo Alto, CA, USA
[2]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA

Abstract (250/250 words)

Selective publication and reporting in individual papers compromise the scientific record, but are meta-analyses as compromised as their constituent studies? We systematically sampled 63 moderately large meta-analyses (at least 40 studies per meta-analysis) in *PLOS One*, top medical journals, top psychology journals, and Metalab, an online, open-data database of developmental psychology meta-analyses. We empirically estimated publication bias in each. Across all meta-analyses, "statistically significant" results in the expected direction were only 1.20 times more likely to be published than "nonsignificant" results or those in the unexpected direction (95% CI: [0.94, 1.53]), with a confidence interval substantially overlapping the null. Comparable estimates were 0.82 for meta-analyses in *PLOS One*, 1.23 for top medical journals, 1.54 for top psychology journals, and 4.68 for Metalab. We estimated that for 87% of meta-analyses, the amount of publication bias that would be required to attenuate the point estimate to the null exceeded the amount of publication estimated to be actually present in the vast majority of meta-analyses from the relevant scientific discipline (exceeding the 95^{th} percentile of publication bias). Study-level measures ("statistical significance" with a point estimate in the expected direction and point estimate size) did not indicate more publication bias in higher-tier versus lower-tier journals, nor in the earliest studies published on a topic versus later studies. Overall, the mere act of performing a meta-analysis with a large number of studies (at least 40) and that includes non-headline results may largely mitigate publication bias in meta-analyses, suggesting optimism about the validity of meta-analytic results.

Key words: publication bias, selective reporting, meta-analysis, reproducibility, scientific method

Significance Statement (118/120 words)

How much does publication bias compromise the integrity of meta-analysis results? In a systematic, interdisciplinary sample of meta-analyses, we found that "statistically significant" results in the expected direction were only 1.20 times more likely to be published than "nonsignificant" results, and statistical uncertainty was also consistent with no publication bias on average. These estimates ranged from 0.82 to 4.68 for meta-analyses from different sources. For the large majority of meta-analyses, conclusions are robust to plausible amounts of publication bias. There seemed to be no more publication bias in individual studies published in higher-tier journals versus lower-tier journals, nor in the earliest studies published on a topic. Our results suggest optimism about the validity of most meta-analytic estimates.

Publication bias – that is, the selective publication of "statistically significant" results ([1]) – has compromised the integrity of the scientific record ([2]). Empirical results often replicate at lower than expected rates (e.g., ([3]–[6])), "p-hacking" (i.e., intentionally or unintentionally re-running analyses to attain "statistically significant" results) appears widespread ([7], [8]), and results in some top social sciences journals exhibit severe publication bias ([9], [10]). Most attention on publication bias and scientific credibility to date has focused on individual published papers, often those in higher-tier journals. In contrast, meta-analyses represent an arguably higher standard of scientific evidence, and the implications of publication bias in individual papers on meta-analyses are not clear. Are meta-analyses of biased literatures simply "garbage in, garbage out", or are meta-analyses more robust to publication bias than are their constituent studies?

Some existing work has considered publication bias in meta-analyses by using statistical methods related to funnel plot asymmetry (see ([11]) for a review) to estimate the percentage of systematically sampled meta-analyses with "statistically significant" publication bias; these estimates include $7-18\%$ among Cochrane Database meta-analyses ([12]), 13% among meta-analyses in *Psychological Bulletin* and the Cochrane Database ([13]), and 27% among medical meta-analyses ([14]). However, asymmetry-based methods of estimating publication bias have statistical assumptions, including a lack of heterogeneity, that appear to be violated for most meta-analyses ([11], [12], [15]). Furthermore, hypothesis tests of publication bias are underpowered for most meta-analyses ([16]) and do not naturally provide estimates of the actual severity of publication bias itself. Other investigators have reported strong publication bias in meta-analyses by testing for the difference in the number of observed "significant" studies in the meta-analyses to the number expected, assuming the true effect size in each study is equal to the meta-analytic point estimate ([17], [18]). This approach would often overestimate publication bias severity if true effects are heterogeneous ([19]), and indeed there is strong evidence for substantial heterogeneity in many meta-analyses ([20]). Other methods that have been used to empirically assess publication bias often also require homogeneous true effects ([13]).

We built upon prior work by conducting a new meta-analysis of meta-analyses that we systematically collected from four sources: the interdisciplinary journal *PLOS One*, three top medical journals, three top psychology journals, and an online, open-data repository of meta-analyses on developmental psychology called "Metalab" ([21], [22]). These meta-analyses spanned a range of journals and disciplines. We focused on estimating the severity of publication bias, defined here as the relative probability of publication for statistically "significant" results with point estimates in the expected direction versus for statistically "nonsignificant" results or results in the unexpected direction ([23]). This operationalization of publication bias is intended to provide an intuitively tractable estimate of the actual severity of publication bias and to accommodate a realistic form of publication bias that favors "statistically significant" results. Empirical evidence suggests this model of publication

bias is often more realistic than that underlying asymmetry tests, which assume publication bias favors large point estimates rather than "significant" p-values and does not affect very large studies (7, 24). We conducted a meta-meta-analysis to arrive at overall, within-group, and within-discipline estimates of publication bias severity. Additionally, to explore hypothesized study-level contributors to publication bias, we assessed whether studies published in higher-tier journals exhibit more publication bias than those in lower-tier journals (25, 26) and whether the chronologically first few studies published on a topic exhibit more publication bias than later studies (the "Proteus effect"; (27)). To do so, we assessed the associations of journal tier and study chronology with two study-level indicators of publication bias: first, "statistical significance" with a point estimate in the expected direction, and second, point estimate size.

1. Methods

1.1. Systematic search methods

The methods and analytic decisions were preregistered (`https://osf.io/qr9h8/`); the Supplement describes and justifies deviations from this protocol. Extensive documentation on each step of the search process, data extraction, data cleaning, and analysis is available online (`https://osf.io/cz8tr/`). We systematically searched for meta-analyses from four sources: (1) *PLOS One*; (2) four top medical journals:[1] *New England Journal of Medicine*, *Journal of the American Medical Association*, *Annals of Internal Medicine*, and *Lancet*; (3) three top psychology journals: *Psychological Bulletin*, *Psychological Science*, and *Perspectives on Psychological Science*; and (4) Metalab, an online, unpublished repository of meta-analyses on developmental psychology. Metalab is a database of meta-analyses on developmental psychology whose datasets are made publicly available and are continuously updated; these meta-analyses are often released online prior to publication in peer-reviewed journals (21, 22). We selected these sources in order to represent a range of disciplines, particularly via the inclusion of *PLOS One* meta-analyses. Additionally, because selection pressures on meta-analyses themselves may differ by journal tier, we chose sources representing higher-tier journals, a middle-tier journal with an explicit focus on publishing all methodologically sound papers regardless of results (*PLOS One*), and a source that is not a standard peer-reviewed journal (Metalab).

For the three published sources, we reverse-chronologically reviewed each meta-analysis published after 2013 until we had obtained data suitable for re-analysis to fulfill or surpass prespecified sample sizes (Supplement). Our inclusion criteria were: (1) the meta-analysis comprised at least 40 studies to enable reasonable power and asymptotic properties to estimate publication bias (23); (2) the meta-analyzed studies tested a hypothesis (e.g., they were not purely descriptive); and (3) we could obtain study-level point estimates and standard errors as described in Section 1.2. For *PLOS One*, we defined three disciplinary categories (social sciences, natural sciences, and medicine) and searched until we had obtained at least 10 usable meta-analytic estimates per discipline. Because of the relatively few meta-analyses published in the top medical and top psychology journals, we included all eligible meta-analyses published after 2013.[2] For the unpublished source, Metalab, we used publicly available data to include the meta-analyses (28–32) meeting the above inclusion criteria. We conducted the searches on 2018-12-20 (*PLOS One*), 2019-5-13 (the top medical journals), 2019-5-4

[1] Ultimately, no meta-analyses in *New England Journal of Medicine* met inclusion criteria, so this journal was not represented in analyses.

[2] We pre-specified that we would search these sources until we reached 20 medical and 20 psychology meta-analyses, but anticipating correctly that fewer than 20 would actually have been published in the specified time frame.

(the top psychology journals), and 2019-5-26 (Metalab). For *PLOS One*, we used PubMed to search "meta analysis[Title] AND 'plos one'[Journal]", restricting the search to 2013 onward. For the top medical and top psychology journals, we either used comparable PubMed search strings provided online (https://osf.io/cz8tr/) or we directly searched the journal's website for papers with "meta-analysis" in the title or abstract. For Metalab, we used Table 1 from (33) to screen 10 existing Metalab meta-analyses using our inclusion criterion for the number of point estimates.

1.2. Data extraction

We extracted study-level data using publicly available datasets, datasets we obtained by contacting authors, or data we manually extracted from published forest plots or tables. We also excluded any unpublished studies, such as dissertations, book chapters, and estimates that the meta-analysts obtained by contacting other investigators, because these studies may be subject to different selection pressures from those affecting published studies. To minimize data entry errors, we used independent dual coding by a team of research assistants (Acknowledgments) and the first author, and we used stringent quality checks to verify data entry. Details of the data extraction process appear in the Supplement, and the final corpus of meta-analyses is publicly available (excluding those for which we could obtain data only by contacting the authors) and is documented for use in future research (https://osf.io/cz8tr/).

For each meta-analysis in the top medical and top psychology groups, we coded each meta-analyzed study by journal, publication year, and the journal's Scimago impact rating (34). Scimago ratings are conceptually similar to impact factors, but weight a journal's citations by the impact of the citing articles rather than treating all citations equally. Additionally, unlike impact factors, Scimago ratings are available in a single, standardized online database (34). We coded each study by its journal's Scimago rating in 2019 or the most recent available rating regardless of the study's publication year in order to avoid conflating overall secular trends in scientific citations with relative journal rankings. We defined "higher-tier" journals as those surpassing a Scimago rating of 3.09 for psychology (chosen such that the lowest-ranked "higher-tier" journal was *Journal of Experimental Psychology: General* and all specialty journals were considered "lower-tier") or 7.33 for medicine (such that the lowest-ranked "higher-tier" journal was *Annals of Internal Medicine*)[3]. All other journals were defined as "lower-tier".

To assess whether publication bias was more severe for the first few studies published on a topic compared to later studies, we coded studies as being published "early" versus "later" as follows. For each meta-analysis, we considered the first chronological year in which any study was published; if multiple studies were published that year, then all point estimates from those studies were coded as "early". If instead only one study was published during the first year, then all point estimates from all studies published during the chronologically first *two* years were coded as "early". All point estimates not coded as "early" were coded as "later".

[3]We set these thresholds based on the discipline of the meta-analysis' journal, not that of the study's journal, because we did not have fine-grained data on each study's disciplinary category. Therefore, in principle, a study published in a medical journal but included in a psychology meta-analysis might be spuriously coded as "higher-tier" because it was compared to the lower threshold for psychology. However, the impact on analysis would likely be minimal. Of the 84% of unique journals in our dataset that were included in journal tier analyses and that also had a topic categorization available in the Scimago database, only three journals with the string "medic*" in the Scimago categorization were published in psychology meta-analyses, and manual review indicated these journals were genuinely interdisciplinary rather than purely medical. Additionally, these journals would have been coded as "lower-tier" regardless of which threshold was applied. No journals with "psych*" in the Scimago categorization were included in medical meta-analyses.

1.3. Primary statistical analyses

1.3.1 Estimates of publication bias severity

We estimated publication bias using selection models, a class of statistical methods that assume that publication bias selects for studies with statistically "significant" results in the expected direction, such that these results (which we term "affirmative") are more likely to be published than statistically "nonsignificant" results or results in the unexpected direction (which we term "nonaffirmative") by an unknown ratio. This selection ratio represents the severity of publication bias: for example, a ratio of 30 would indicate severe publication bias in which affirmative results are 30 times more likely to be published than nonaffirmative results, whereas a ratio of 1 would indicate no publication bias, in which affirmative results are no more likely to be published than nonaffirmative results. Selection models essentially detect the presence of nonaffirmative results arising from analyses that were conducted but not reported; these results are therefore "missing" from the published and meta-analyzed studies. Specifically, we used a selection model that specifies a normal distribution for the true effect sizes, weights each study's contribution to the likelihood by its inverse-probability of publication based on its affirmative or nonaffirmative status, and uses maximum likelihood to estimate the selection ratio (23). We used a selection model approach to allow for a realistic mechanism of publication bias that operates based on "statistical significance" and to accommodate heterogeneous true effects (see (11) and (35) for reviews).

As in standard meta-analysis, selection models assume that studies' point estimates are independent, but this assumption may be violated when some studies contribute multiple point estimates to a meta-analysis (e.g., estimates of a single intervention's effect on different subject populations). To minimize the possibility of non-independence, we randomly selected one point estimate per study within each meta-analysis and then fit the selection model to only these independent estimates. Because the "expected" effect direction differed across meta-analyses, we first synchronized the signs of all point estimates so that positive effects represented the expected effect direction. To this end, we first re-analyzed all point estimates using restricted maximimum likelihood estimation and the R package `metafor` and, treating the sign of the resulting pooled point estimate as the expected effect direction, reversed the sign of all point estimates for any meta-analysis with a negative pooled point estimate. We fit a selection model to estimate the inverse of the selection ratio and its standard error (23). We then used robust methods (36) to meta-analyze the log-transformed estimates of the selection ratio, approximating their variances via the delta method.

To characterize the upper limit of publication bias that might be expected in our sample of meta-analyses, we calculated the maximum estimate of the selection ratio; however, this is a crude, upward-biased measure because sampling error introduces more variation in the study-level estimates than in the underlying true effects (37). Therefore, we additionally estimated the 95^{th} quantile of the true selection ratios using a nonparametric shrinkage method that accounts for sampling error (37). We conducted these analyses across all meta-analyses as well as by group and, within the *PLOS One* group, by discipline. We conducted a number of sensitivity analyses to assess the impacts of possible violations of modeling assumptions, all of which yielded similar results (Supplement).

1.3.2 Study-level indicators of publication bias

For the top medical and top psychology meta-analyses, but not those in *PLOS One* or Metalab[4], we assessed the association of the tier of the individual study's journal with two study-level measures of publication bias: whether the study was affirmative[5] per Section 1.3 and the size of its point estimate. To characterize the size of each study's point estimate relative to those of other studies on the same topic, we computed within-meta-analysis percentiles of point estimates. We used percentiles rather than raw effect sizes to provide a metric that is comparable across meta-analyses regardless of their differing numbers of studies, mean effect sizes, and measures of effect size. We estimated the percentages of affirmative results and mean point estimate percentiles by journal tier (higher-tier vs. lower-tier), and by study chronology (early vs. later publication date). As a post hoc analysis, we estimated the overall risk ratio of an affirmative result comparing higher-tier to lower-tier journals (i.e., the relative probability of an affirmative result in higher-tier versus lower-tier journals) using log-binomial generalized estimating equations (GEE) models with robust inference to account for correlation of point estimates within studies and meta-analyses (38, 39). We also conducted a comparable set of descriptive and regression analyses regarding a study's chronology, including all four groups of meta-analyses.

2. Primary Results

2.1. Corpus of meta-analyses

Overall, our dataset comprised 63 meta-analyses: 33 in *PLOS One*, 7 in top medical journals, 18 in top psychology journals, and 5 in Metalab. A spreadsheet describing the scientific topics of each meta-analysis and our methods of data extraction for each is available online (`https://osf.io/cz8tr/`). Of the *PLOS One* meta-analyses, 10 were categorized as medical, 11 were social sciences, and 12 were natural sciences. We obtained study-level data from publicly available datasets for 27 meta-analyses, by scraping published figures or tables for 23 meta-analyses, and by contacting authors for the remaining 13 meta-analyses. The total number of point estimates after the removal of unpublished studies was 12494, and the meta-analyses comprised a median of $n = 80$ point estimates each. The total numbers of point estimates within each group are provided in Tables 2 and 3. When we re-analyzed the published studies within each meta-analysis using robust meta-analysis to accommodate clustering of point estimates within studies (36), the mean magnitude of pooled point estimates after synchronizing their directions as described in Section 1.3 and without correction for publication bias was 0.52 for standardized mean differences ($n = 29$), 1.05 for ratio measures, including odds ratios, hazard ratios, risk ratios, and mean ratios ($n = 13$), and 0.22 for Pearson's correlations ($n = 15$). An additional 6 meta-analyses used other, less common types of effect size.

Among top medical and top psychology meta-analyses (those used in journal tier analyses), 18% of point estimates were published in higher-tier journals. Among these meta-analyses that were published in top medical journals, 4% of estimates were in higher-tier journals. Among the meta-analyses in top psychology journals, 18% of estimates were in higher-tier journals. We extracted journal tier data for 95% of point estimates in top medical and top psychology meta-analyses; some

[4] As preregistered, we excluded *PLOS One* because the meta-analyses' highly diverse topics and subdisciplines made it prohibitively challenging to define journal tier thresholds that would be reasonable for all meta-analyses. We excluded Metalab because our pilot work suggested that almost none of the meta-analyzed studies were published in higher-tier journals.

[5] We conducted sensitivity analyses in which we instead considered two-tailed statistical "significance" regardless of the estimate's sign, which yielded similar results and are described in Section 2.3.

data were missing because the study's journal had apparently not received a Scimago ranking, and we excluded these point estimates in journal tier analyses. We manually coded journal year data for a convenience sample of all meta-analyses, including 75% of all point estimates; 3% of these point estimates were published early (ranging from 3% to 5% within the four groups).

2.2. Estimates of publication bias severity

We estimated the selection ratio using a total of 55 meta-analyses; we excluded estimates from meta-analyses with fewer than 3 affirmative studies or fewer than 3 nonaffirmative studies to minimize problems of statistical instability ([23]). The analyzed meta-analyses had a median of $n = 49$ independent point estimates per meta-analysis, with an overall total of 3781 estimates. Via meta-meta-analysis, we estimated that affirmative results were 1.20 times more likely to be published than non-affirmative results (95% CI: [0.94, 1.53]). The estimated 95^{th} quantile of the true selection ratios was 3.72. Table 1 and Figure 2 display estimates by disciplinary group and by individual meta-analysis, respectively. In *PLOS One* meta-analyses, affirmative results were an estimated 0.82 (95% CI: [0.61, 1.12]) times as likely to be published than nonaffirmative results, which is in fact in the direction opposite what would be expected with publication bias favoring affirmative results (albeit with a wide confidence interval that overlaps the null). Meta-analyses in top medical journals (selection ratio estimate: 1.23; 95% CI: [0.50, 3.04]; $p = 0.31$ vs. *PLOS One*) and top psychology journals (estimate: 1.54; 95% CI: [1.02, 2.34]; $p = 0.01$ vs. *PLOS One*) exhibited relatively mild publication bias in the expected direction. In contrast, in Metalab, affirmative results were an estimated 4.68 times more likely to be published than non-affirmative results (95% CI: [1.95, 11.21]; $p = 0.005$ vs. *PLOS One*), though the wide confidence interval indicated considerable uncertainty.

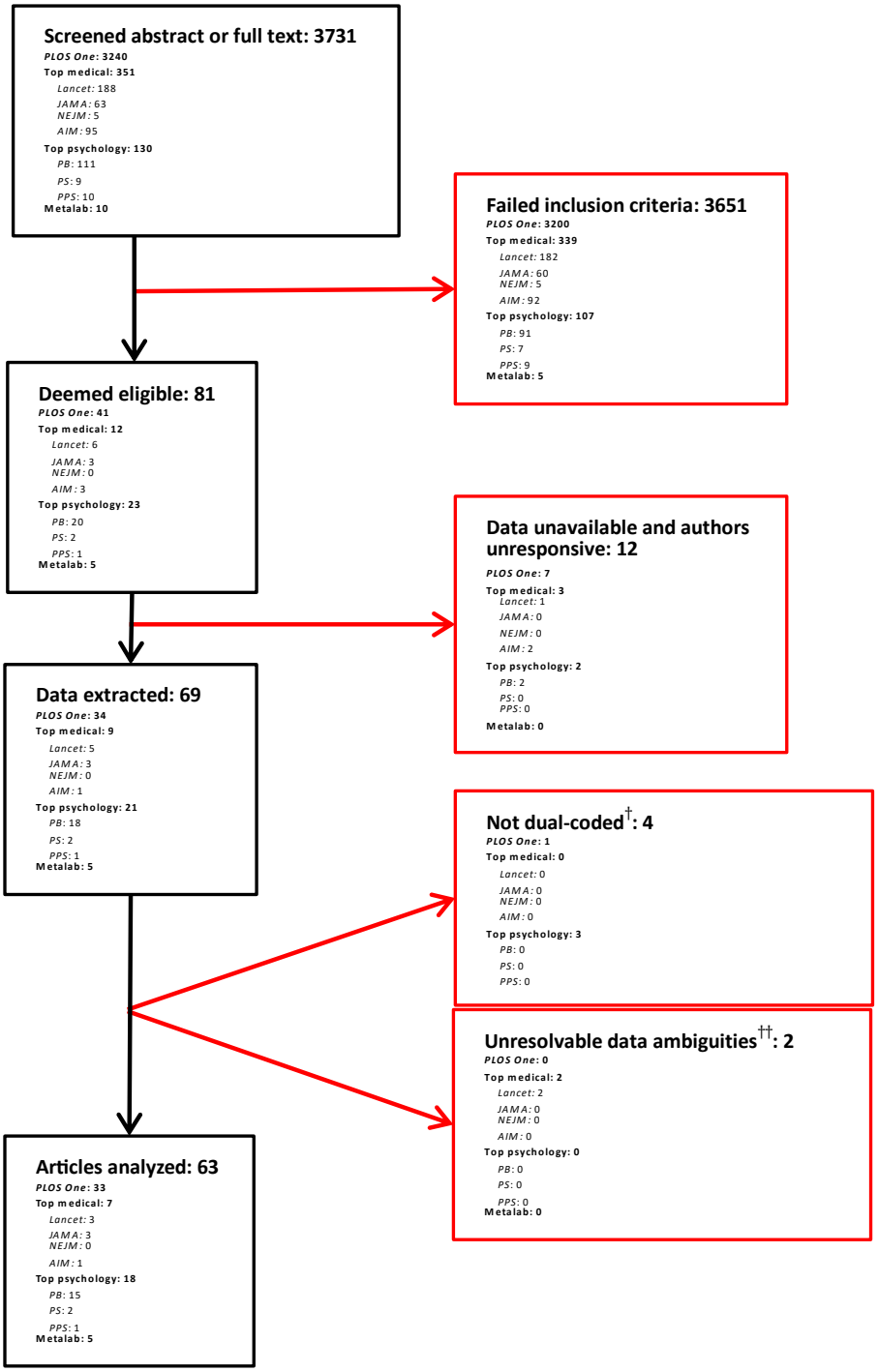

Figure 1: *PRISMA flowchart depicting article screening and exclusion process. Black boxes on the left indicate meta-analyses that remained in the pool of assessed articles at each step; red boxes on the right indicate meta-analyses excluded at each step of assessment. †: Meta-analyses not dual-coded due to data collection time constraints. Metalab is omitted because these meta-analyses were intentionally single-coded. ††: Meta-analyses that were excluded due to unresovable ambiguities in their datasets; see Supplement for details.*

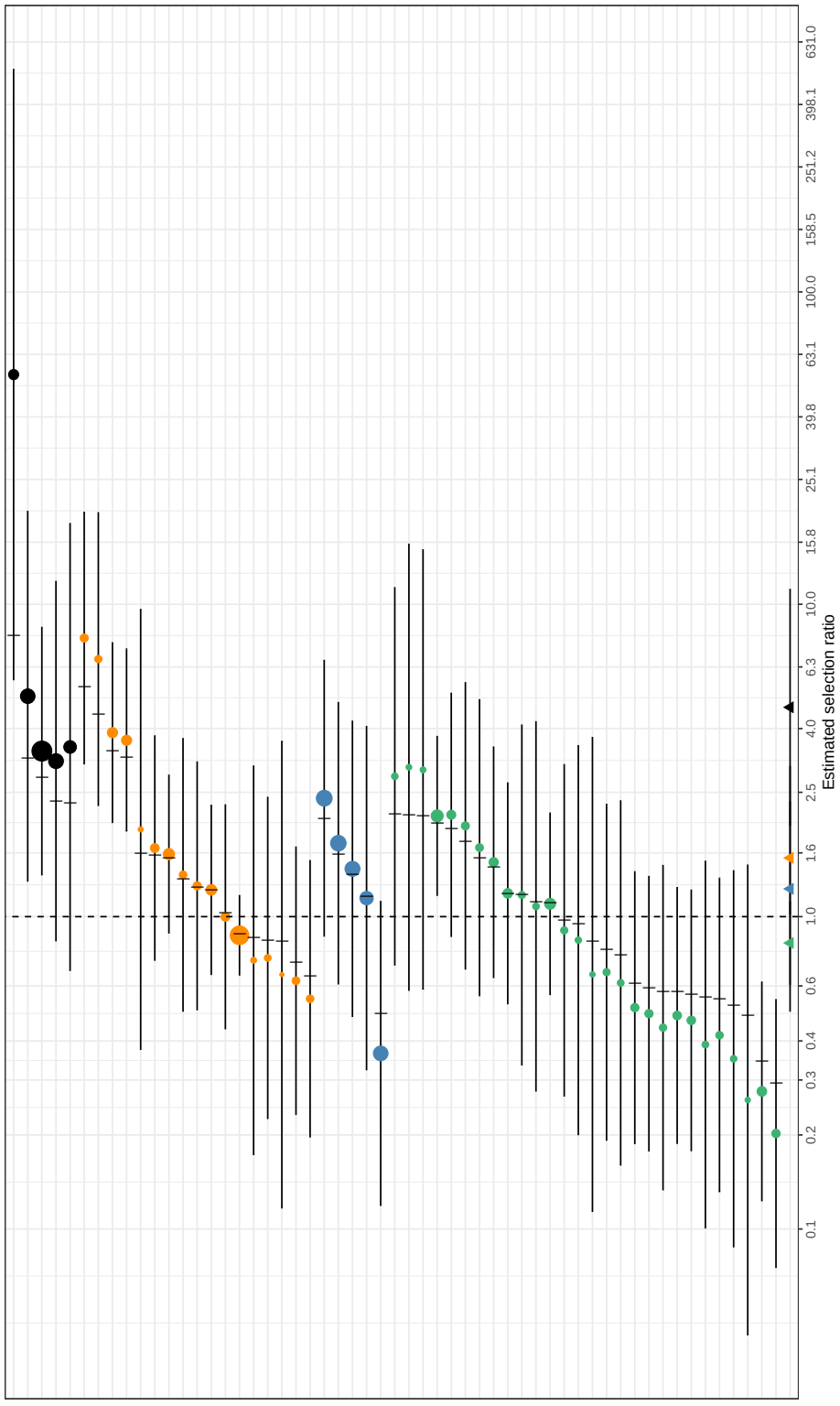

Figure 2: *Selection ratio estimate for each meta-analysis, ordered by group and by the estimate of the meta-analysis' true selection ratio (vertical tick marks). Colored circles represent point estimates of the selection ratio in each meta-analysis, with areas proportional to the meta-analysis' relative weight in the within-group meta-analyses of selection ratios. The x-axis is presented on the log scale. Error bars represent 95% confidence intervals. The vertical dashed line represents the null (no publication bias).*

2.3. Percentages of affirmative results

Across all four groups of meta-analyses, 50% of point estimates (95% CI: [48%, 52%])[6] were affirmative, and 55% of point estimates (95% CI: [53%, 57%]) were "significant" regardless of point estimate sign. Regarding journal tier, the percentage of affirmative results in top medical and top psychology meta-analyses ($n = 7622$ point estimates) was nearly identical for higher-tier journals (56, 95% CI: [49, 62]) and lower-tier journals (58, 95% CI: [54, 61]); see Table 2. However, among the smaller sample of point estimates from top medical meta-analyses alone ($n = 483$), only 26% of results in higher-tier journals were affirmative, compared to 41% in lower-tier journals. Overall, results in higher-tier journals were an estimated 1.04 times as likely to be affirmative as those in lower-tier journals (95% CI: [0.98, 1.10]; $p = 0.16$).

Regarding studies' chronological ordering, the percentage of affirmative results was almost exactly the same for early results (54%; 95% CI: [41%, 66%]) as for later results (54%; 95% CI: [52%, 57%]), though this pattern appeared to vary somewhat across the four major groups of meta-analyses (see final two columns of Table 2). Overall, early results were an estimated 0.90 times as likely to be affirmative than later results (95% CI: [0.76, 1.05]; $p = 0.19$). Risk ratio estimates for each meta-analysis for both journal tier and study chronology are presented in Supplementary Figures S2-S3. We conducted sensitivity analyses in which we considered publication in terms of two-tailed "significance" (i.e., a two-tailed $p < 0.05$ regardless of point estimate sign) rather than "affirmative" status. Similarly to primary analyses, this sensitivity analysis estimated that higher-tier journals were 0.90 (95% CI: [0.76, 1.05]; $p = 0.19$) times as likely to be "significant" as results in lower-tier journals and estimated that early results were 0.94 (95% CI: [0.81, 1.08]; $p = 0.38$) times as likely to be "significant" as later results. We also conducted a sensitivity analysis in which we excluded from the "higher-tier" designation a single journal (*Journal of Educational Psychology*) that had contributed 47% of the higher-tier point estimates. After excluding this journal, higher-tier point estimates appeared less likely than lower-tier point estimates to be affirmative (Supplement).

Group	k	$\widehat{SR}$ [95% CI]	max $\widehat{SR}$	q_{95}	$\widehat{\tau}$	p-value vs. *PLOS*
Overall	55	1.20 [0.94, 1.53]	54.37	3.72	0.66	–
PLOS One	28	0.82 [0.61, 1.12]	3.01	1.73	0.55	Ref.
Medical	7	0.94 [0.47, 1.90]	3.01	1.62	0.43	
Natural sciences	11	0.59 [0.32, 1.06]	2.81	1.55	0.63	
Social sciences	10	1.07 [0.66, 1.73]	2.95	1.75	0.40	
Top medical	5	1.23 [0.50, 3.04]	2.39	1.83	0.45	0.31
Top psychology	17	1.54 [1.02, 2.34]	7.80	4.84	0.63	0.01
Metalab	5	4.68 [1.95, 11.21]	54.37	9.75	0.43	0.005

Table 1: *Overall and within-group estimates of the selection ratio ($\widehat{SR}$) from robust meta-analyses. k: number of analyzed meta-analyses; max $\widehat{SR}$: maximum estimated selection ratio among the group's meta-analyses; q_{95}: estimated 95th quantile of true selection ratios among the group's meta-analyses; $\widehat{\tau}$: meta-analytic estimate of the standard deviation of log-selection ratios; p-value: meta-regressive inference for the difference in publication bias severity vs. all* PLOS One *meta-analyses.*

[6] Confidence intervals for all descriptive statistics estimated at the individual study level use cluster-robust inference with clustering by meta-analyzed study (38, 39).

Group	n	P(affirm)	P(affirm | top-tier)	P(affirm | lower-tier)	P(affirm | early)	P(affirm | later)
PLOS One	3636	0.35 [0.32, 0.38]			0.46 [0.22, 0.69]	0.30 [0.26, 0.34]
Top medical	558	0.37 [0.26, 0.48]	0.26 [0.06, 0.47]	0.41 [0.29, 0.53]	0.54 [0, 1]	0.37 [0.26, 0.48]
Top psychology	7501	0.59 [0.56, 0.62]	0.56 [0.50, 0.63]	0.59 [0.56, 0.63]	0.53 [0.37, 0.69]	0.59 [0.56, 0.62]
Metalab	799	0.43 [0.38, 0.48]	0.53 [0.29, 0.76]	0.41 [0.36, 0.46]	0.64 [0.37, 0.91]	0.41 [0.36, 0.46]

Table 2: *Probabilities of affirmative results overall, by journal tier, and by a study's chronology as one of the first few published ("early") versus as one of the later studies published ("later").* PLOS One *was omitted from journal tier analyses. Cluster-robust confidence intervals are presented, accounting for correlation of p-values within studies. n: number of point estimates in group (which may exceeded number in each analysis due to missing data); P(affirm): probability of an affirmative result.*

Group	n	$\bar{Q}_{\text{higher-tier}}$	$\bar{Q}_{\text{lower-tier}}$	$\bar{Q}_{\text{early}}$	$\bar{Q}_{\text{later}}$
PLOS One	3636			0.47 [0.36, 0.59]	0.51 [0.49, 0.54]
Top medical	558	0.51 [0.35, 0.67]	0.51 [0.45, 0.57]	0.34 [0.10, 0.57]	0.52 [0.47, 0.57]
Top psychology	7501	0.50 [0.47, 0.54]	0.51 [0.49, 0.52]	0.56 [0.44, 0.69]	0.51 [0.49, 0.52]
Metalab	799	0.61 [0.52, 0.71]	0.49 [0.46, 0.52]	0.62 [0.43, 0.80]	0.49 [0.46, 0.52]

Table 3: *Mean within-meta-analysis percentiles of point estimates ($\bar{Q}$) overall, by journal tier, and by a study's status as one of the first three ("early") published versus as one of the later studies published.* PLOS One *was omitted from journal tier analyses. n: number of point estimates in group (which may exceeded number analyzed due to missing data).*

2.4. Size of point estimates

Combining all four groups of meta-analyses, the mean within-meta-analysis percentile of point estimates in higher-tier journals (0.51; 95% CI: [0.48, 0.54]) was identical to that in lower-tier journals (0.51; 95% CI: [0.49, 0.52]). There was almost no difference in mean percentiles comparing studies in higher- versus lower-tier journals (estimate: 0; 95% CI: [−0.01, 0.02]; $p = 0.21$). Within groups, results were mixed, with meta-analyses from top medical journals perhaps showing somewhat smaller point estimates in early studies, Metalab showing the opposite pattern, and the remaining two groups showing little difference (Table 3, final two columns). Considering studies' chronological ordering, the mean percentile in early studies (0.54; 95% CI: [0.45, 0.64]) was also similar to that in later studies (0.51; 95% CI: [0.49, 0.52]); the estimated difference was 0.03 (95% CI: [−0.02, 0.07]; $p = 0.29$).

3. Exploratory Results

3.1. Sensitivity to varying amounts of publication bias

As an alternative method of considering the possible impact of publication bias on meta-analysis results, we conducted post hoc sensitivity analyses to assess the severity of publication bias that would be required to "explain away" the results of each meta-analysis (40), rather than to estimate the amount of publication bias in each meta-analysis as we did in the main analyses. The sensitivity analysis methods assess: (1) the minimum selection ratio that would be required to attenuate a meta-analytic pooled point estimate to the null; and (2) the minimum selection ratio that would be required to shift the confidence interval to include the null. They also allow estimation of a

"worst-case" pooled point estimate and confidence interval under maximal publication bias in which affirmative studies are almost infinitely more likely to be published than nonaffirmative studies; these worst-case estimates are obtained by simply meta-analyzing only the nonaffirmative studies (40). These methods obviate the distributional and independence assumptions required for our main analysis models, providing a form of sensitivity analysis for the main results.

For these analyses, we retained all point estimates from each meta-analysis and used a robust sensitivity analysis model to account for clustering and non-normality (40). Worst-case pooled point estimates remained in the same direction as the pooled point estimate for 58% of meta-analyses, indicating that no amount of publication bias under the assumed model would suffice to shift the point estimate to the null for this majority of meta-analysis. Among these meta-analyses, the worst-case point estimate was on average 26% as large as the pooled point estimate. Considering all meta-analyses, the worst-case 95% confidence interval limit excluded the null for 23% of meta-analyses, indicating that no amount of publication bias under the assumed model would suffice to shift the confidence interval to include the null. The estimated 5^{th} and 10^{th} percentiles of the true selection ratios indicated that for 95% of meta-analyses, affirmative results would need to be at least 1.48 times more likely to be published than nonaffirmative results in order to attenuate the pooled point estimate to the null; and for 90% of meta-analysis, this ratio would need to be at least 3.19. In fact, for 89% of meta-analyses, the amount of publication bias required to attenuate the pooled point estimate to the null exceeded our previous empirical estimate of the actual amount of publication bias in 95% of meta-analyses from the relevant group (c.f. Table 1, column "q_{95}"). Additionally, for 73% of meta-analyses, the amount of publication bias required to shift the confidence interval to include the null exceeded this 95^{th} percentile empirical estimate of actual publication bias severity[7].

3.2. Effect of including unpublished studies in Metalab

As discussed above, publication bias appeared more severe in Metalab than in the published sources of meta-analyses, though the small sample size in Metalab precludes strong conclusions. We speculate that particularly small sample sizes in developmental psychology research (averaging 18 subjects; (21)) may contribute to publication bias in this group. We additionally investigated the effect of including unpublished studies on publication bias estimates in Metalab; unpublished studies constituted on average 14% of the independent point estimates included in the selection models. When fit to datasets that include unpublished studies, selection models detect the presence of nonaffirmative results arising from analyses that were conducted, but not reported in any published *or* unpublished source that was available for inclusion in the meta-analysis. Across the five meta-analyses, estimates of publication bias typically *increased* upon inclusion of the unpublished studies; selection ratios increased by on average 1.96-fold, and the ratios of change ranged from 0.76-fold to 4.76-fold. This is consistent with previous findings suggesting that the inclusion of unpublished studies did not consistently reduce publication bias in these meta-analyses (33).

3.3. Selection ratios for studies in higher- vs. lower-tier journals

Our primary analyses regarding journal tiers considered study-level indicators of publication bias (i.e., percentages of affirmative results and sizes of point estimates), rather than estimates of the selection ratio by journal tier, because we anticipated correctly that very few meta-analyses would have enough affirmative and nonaffirmative results within each journal tier category to permit such

[7] Among meta-analyses for which the confidence interval did not already include the null, this percentage increased slightly to 83%.

an analysis. However, we did conduct this analysis for the two meta-analyses that had, within each journal tier category, at least 40 independent point estimates, including at least 3 affirmative and 3 nonaffirmative results. Both meta-analyses were in the top psychology group and were published in *Psychological Bulletin*. For the first (41), the estimated selection ratio among the 71 lower-tier estimates was 3.77 (95% CI: [1.49, 9.54]) and among the 47 higher-tier estimates was 3.19 (95% CI: [1.17, 8.74]). For the second (42), the estimated selection ratio among the 528 lower-tier estimates was 0.76 (95% CI: [0.56, 1.05]) and among the 64 higher-tier estimates was 5.92 (95% CI: [1.59, 22.04]). Thus, in the first meta-analysis, given the fairly wide confidence intervals, there appeared heuristically to be little difference between journal tiers in publication bias severity. However, in the second, the publication bias appeared considerably more severe in the higher-tier journals.

These differing results might reflect heterogeneity in publication bias severity across individual journals, even within each journal tier category. In both meta-analyses, a majority of higher-tier results were published in just one or two journals. For the first meta-analysis, 66% of higher-tier results were published in *Journal of Consumer Research* and *Journal of Personality and Social Psychology* combined. For the second, 64% of higher-tier results were published in *Journal of Educational Psychology*, the same journal whose exclusion in an aforementioned sensitivity analysis had reduced the estimated risk ratio of an affirmative result in higher- versus lower-tier journals from 1.04 to 0.83 (Section 2.3). However, these results are merely exploratory and do not allow us to parse potential differences in publication bias across individual journals into effects of, for example, the journals' subdisciplines, their editorial practices, and authors' submission practices.

4. Discussion

Our systematic analysis of meta-analyses spanning several disciplines suggested that publication bias is perhaps milder than expected in meta-analyses published in *PLOS One*, top medical journals, and top psychology journals. Study-level measures of publication bias, namely the percentage of affirmative results and the size of point estimates, indicated that publication bias did not differ meaningfully for original studies published in higher-tier versus lower-tier journals, nor for the first few studies published on a topic versus for later studies. Secondary analyses that assessed the sensitivity of meta-analyses' findings to varying amounts of hypothetical publication bias, rather than estimating the amount of publication bias itself, corroborated primary findings and suggested that the major conclusions of most meta-analyses are robust to plausible amounts of publication bias. We excluded unpublished studies from all meta-analyses; publication bias might have been yet milder had we included these results, though exploratory findings in Metalab casted some doubt on this possibility.

Our estimates of publication bias were lower than we expected. For comparison, previous work examining social sciences lab experiments estimated selection ratios from 10 to 48 (Tables 1 and 2 in (9)), which are an order of magnitude larger than our estimates. Others have estimated publication bias by prospectively or retrospectively following cohorts of study protocols submitted to specific ethics committees or funded by specific granting agencies. In a systematic review of such cohort studies (typically within the medical domain), eight estimates of parameters qualitatively similar to the selection ratio ranged from approximately 0.73 to 3.51 (Table 5 in (43)))[8]. A cohort study published since that review followed social sciences experiments funded through a certain granting agency, estimating a selection ratio of approximately 2.95 (45). Our research question and

[8]Some estimates were reported on the odds ratio scale. To put these estimates on a scale comparable to a selection ratio, which is essentially a risk ratio, we used a square-root approximation that does not rely on the rare-outcome assumption (44).

methodology differed in an important manner from those of these previous studies: rather than estimating publication bias in original studies themselves, we estimated publication bias in the published results that were included in meta-analyses. It is plausible that meta-analyzed results exhibited relatively little publication bias compared to original papers' results because meta-analyses deliberately attempt to include all results on a topic, including replication studies and null results published in lower-tier journals. However, casting doubt on these explanations, we also found little evidence of increased publication bias in higher-tier journals or in early studies.

We instead speculate that the key alleviator of publication bias in meta-analyses is their inclusion of "non-headline" results, by which we mean results that are reported in published papers but that are de-emphasized (e.g., reported only in secondary or supplemental analyses) and those that meta-analysts obtain through manual calculation or by contacting authors. In contrast, we describe as "headline" results those that are particularly emphasized in published (e.g., those that are included in abstracts or otherwise treated as primary). For comparison, among a semi-systematic sample of headline results[9] ($n = 100$) from studies published in three top psychology journals, 97% were "statistically significant" regardless of the sign of the point estimate (3), compared to only 54% of results from the same three journals in our own corpus but that had been included in some meta-analysis ($n = 238$). Similarly, among headline results (i.e., p-values reported in the abstracts) sampled from papers in four top medical and one top epidemiology journals (46), 78% were "significant" ($n = 15653$), which appears higher than our 50% for all results in meta-analyses in top medical journals ($n = 558$) and our 32% for *PLOS One* meta-analyses on medical topics ($n = 576$).[10] About 90% of headline findings in both medicine and psychology papers were qualitatively described as supporting the investigated hypothesis (47), an estimate that again appears considerably higher than our own. Considering instead non-headline results, the percentages of "significant" p-values in top psychology journals (48) and in an interdisciplinary corpus (49) were 64% and 57% respectively[11]. Among all results in our own meta-analysis corpus, 55% were "significant", which is much closer to the estimates in non-headline results than to estimates in headline results. Holistically, these findings provide preliminary support for the possibility that meta-analyses mitigate publication bias largely through their inclusion of non-headline results, which may be less prone to publication pressures than are headline results.

We sampled meta-analyses across disciplines and journals, yielding findings that we believe generalize to a fairly diverse range of meta-analyses and scientific topics. Nevertheless, we restricted our sample to large meta-analyses (those with at least 40 point estimates in original analyses) for statistical reasons described in Section 1.3. It is plausible that bias could operate differently in large meta-analyses, which might be conducted on well-established rather than nascent literatures (e.g., (52)) or which might have used particularly exhaustive search strategies. Last, although our corpus of meta-analyses represented a wide range of scientific topics, it did not represent topics that are not amenable to meta-analysis (for example, because they use qualitative methods or use statistical methods that do not readily yield simple study-level point estimates). Publication bias may operate differently in such realms of scientific inquiry.

Our research has some limitations. The relatively small number of meta-analyses within each

[9] Specifically, (3) was a replication project that by default selected the key result of the final study of each paper, though for some papers, a different key result was selected.

[10] Our corpus contained only 20 p-values from those four specific journals, but of this small sample, only 35% of results were "significant".

[11] (48) sampled all reported t-statistics in a sample of papers published in 18 prominent psychology and neuroscience journals. (49) aggregated p-values from corpuses that sampled p-values from the bodies of papers in prominent economics journals (50), from papers listed in PubMed (51), and from the Results sections of all open-access articles in PubMed (8).

group precludes strong conclusions about differences in publication bias across the groups. As in all analyses of publication bias, our estimates of publication bias relied on statistical assumptions, namely that true effect sizes are approximately normal prior to selection due to of publication bias, that point estimates are independent, and that publication bias favors affirmative results over nonaffirmative results. To determine which results were "affirmative", we assigned a positive sign to point estimates that agreed in direction with a naïve meta-analytic pooled point estimate that was not corrected for publication bias; this approach effectively assumes that publication bias favors effects in the majority direction. However, we also conducted analyses using sensitivity analysis techniques that obviated the assumptions regarding normality and independence; these findings heuristically corroborated the primary results (Section 3). Additional analyses suggested that the final assumption regarding the mechanism of publication bias was plausible (Supplement).

Overall, our results suggest relatively mild publication bias in meta-analyses in the interdisciplinary journal *PLOS One* and in top psychology and medical journals. Publication bias may in fact have been more severe in an unpublished corpus of developmental psychology meta-analyses, though the sample size was small. Our results suggest that the primary drivers of publication bias in meta-analyses are neither the publication process itself, nor the pressures of publishing individual studies or meta-analyses in higher-tier journals, nor the pressure to publish one of the first studies on a topic. The prioritization of findings within published papers as headline versus non-headline results may contribute more to publication bias than these influences. Thus, the mere act of performing a high-quality meta-analysis that includes non-headline results may itself largely mitigate publication bias, suggesting optimism about the validity of most meta-analytic estimates. Nevertheless, it remains critical to design and analyze meta-analyses with careful attention to publication bias.

Reproducibility

All code, materials, and data aggregated by meta-analysis are publicly available and documented (`https://osf.io/cz8tr/`). Study-level data for each meta-analysis are available for the meta-analyses for which we were able to obtain these data without contacting authors. For the remaining meta-analyses, data cannot be made public at the authors' requests, but they are available upon request by emailing the first author to individuals who have secured permission from the original authors. The preregistered protocol is publicly available (`https://osf.io/qr9h8/`). All post hoc modifications and clarifications to the preregistration are disclosed in the Supplement.

Acknowledgments

Theiss Bendixen, Andrea Lamas-Nino, Maximilian Meier, Leslie Meza, Claire Punturieri, and Yawen Xiang collected and verified data. We thank those meta-analysts who made their datasets publicly available and who responded to our requests for data.

References

[1] Theodore D Sterling. Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association*, 54(285):30–34, 1959.

[2] John PA Ioannidis, Marcus R Munafo, Paolo Fusar-Poli, Brian A Nosek, and Sean P David. Pub-

lication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, 18(5):235–241, 2014.

[3] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[4] Prasad Patil, Roger D Peng, and Jeffrey T Leek. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11(4):539–544, 2016.

[5] Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Juergen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, page 1, 2018.

[6] Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola, Štěpán Bahník, et al. Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

[7] EJ Masicampo and Daniel R Lalande. A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, 65(11):2271–2279, 2012.

[8] Megan L Head, Luke Holman, Rob Lanfear, Andrew T Kahn, and Michael D Jennions. The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3):e1002106, 2015.

[9] Isaiah Andrews and Maximilian Kasy. Identification of and correction for publication bias. Technical report, National Bureau of Economic Research, 2017.

[10] Valen E Johnson, Richard D Payne, Tianying Wang, Alex Asher, and Soutrik Mandal. On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517):1–10, 2017.

[11] Zhi-Chao Jin, Xiao-Hua Zhou, and Jia He. Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine*, 34(2):343–360, 2015.

[12] John PA Ioannidis and Thomas A Trikalinos. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *Canadian Medical Association Journal*, 176(8):1091–1096, 2007.

[13] Robbie CM van Aert, Jelte M Wicherts, and Marcel ALM van Assen. Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLoS One*, 14(4):e0215052, 2019.

[14] Jonathan AC Sterne, David Gavaghan, and Matthias Egger. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology*, 53(11):1119–1129, 2000.

[15] Norma Terrin, Christopher H Schmid, Joseph Lau, and Ingram Olkin. Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, 22(13):2113–2126, 2003.

[16] Petra Macaskill, Stephen D Walter, and Les Irwig. A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, 20(4):641–654, 2001.

[17] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365, 2013.

[18] John PA Ioannidis. Excess significance bias in the literature on brain volume abnormalities. *Archives of General Psychiatry*, 68(8):773–780, 2011.

[19] Valen Johnson and Ying Yuan. Comments on 'an exploratory test for an excess of significant findings' by jpa loannidis and ta trikalinos. *Clinical Trials*, 4(3):254, 2007.

[20] Julian PT Higgins, Simon G Thompson, Jonathan J Deeks, and Douglas G Altman. Measuring inconsistency in meta-analyses. *BMJ*, 327(7414):557–560, 2003.

[21] Christina Bergmann, Sho Tsuji, Page E Piccinini, Molly L Lewis, Mika Braginsky, Michael C Frank, and Alejandrina Cristia. Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, 89(6):1996–2009, 2018.

[22] Molly Lewis, Mika Braginsky, Sho Tsuji, Christina Bergmann, Page Piccinini, Alejandrina Cristia, and Michael C Frank. A quantitative synthesis of early language acquisition using meta-analysis. 2016.

[23] Larry V Hedges. Modeling publication selection effects in meta-analysis. *Statistical Science*, pages 246–255, 1992.

[24] Blakeley B McShane and David Gal. Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, 112(519):885–895, 2017.

[25] Paul A Murtaugh. Journal quality, effect size, and publication bias in meta-analysis. *Ecology*, 83(4):1162–1166, 2002.

[26] Phillipa J Easterbrook, Ramana Gopalan, JA Berlin, and David R Matthews. Publication bias in clinical research. *The Lancet*, 337(8746):867–872, 1991.

[27] Thomas Pfeiffer, Lars Bertram, and John PA Ioannidis. Quantifying selective reporting and the proteus phenomenon for multiple datasets with similar bias. *PLoS One*, 6(3):e18362, 2011.

[28] Hugh Rabagliati, Brock Ferguson, and Casey Lew-Williams. The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science*, 22(1):e12704, 2019.

[29] Katie Von Holzen and Christina Bergmann. A meta-analysis of infants' mispronunciation sensitivity development. In *CogSci: Annual Conference of the Cognitive Science Society. Cognitive Science Society (US). Conference*, volume 2018, page 1157. NIH Public Access, 2018.

[30] Christina Bergmann and Alejandrina Cristia. Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*, 19(6):901–917, 2016.

[31] Alexis Black and Christina Bergmann. Quantifying infants' statistical word segmentation: a meta-analysis. In *39th Annual Meeting of the Cognitive Science Society*, pages 124–129. Cognitive Science Society, 2017.

[32] Angeline Sin Mei Tsui, Krista Byers-Heinlein, and Christopher T Fennell. Associative word learning in infancy: A meta-analysis of the switch task. *Developmental Psychology*, 55(5):934, 2019.

[33] Sho Tsuji, Alejandrina Cristia, Michael C Frank, and Christina Bergmann. Addressing publication bias in meta-analysis: Empirical findings from community-augmented meta-analyses of infant language development. 2019. Preprint retrieved from https://osf.io/preprints/metaarxiv/q5axy/.

[34] Scimago journal and country rank. https://www.scimagojr.com/. Accessed: 2019-07-08.

[35] Blakeley B McShane, Ulf Böckenholt, and Karsten T Hansen. Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11(5):730–749, 2016.

[36] Larry V Hedges, Elizabeth Tipton, and Matthew C Johnson. Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1):39–65, 2010.

[37] Chia-Chun Wang and Wen-Chung Lee. A simple method to estimate prediction intervals and predictive distributions: Summarizing meta-analyses beyond means and confidence intervals. *Research Synthesis Methods*, 2019.

[38] James E Pustejovsky and Elizabeth Tipton. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4):672–683, 2018.

[39] Daniel F McCaffrey and Robert M Bell. Bias reduction in standard errors for linear and generalized linear models with multi-stage samples. In *Proceedings of Statistics Canada Symposium*, pages 1–10, 2002.

[40] MB Mathur and TJ VanderWeele. Sensitivity analysis for publication bias in meta-analyses. 2019. Preprint retrieved from https://osf.io/wdmht/.

[41] Evan Weingarten and J Hutchinson. Does ease mediate the ease-of-retrieval effect? A meta-analysis. *Psychological Bulletin*, 144(3):227, 2018.

[42] Peng Peng, Tengfei Wang, CuiCui Wang, and Xin Lin. A meta-analysis on the relation between fluid intelligence and reading/mathematics: Effects of tasks, age, and social economics status. *Psychological Bulletin*, 145(2):189, 2019.

[43] Kerry Dwan, Carrol Gamble, Paula R Williamson, and Jamie J Kirkham. Systematic review of the empirical evidence of study publication bias and outcome reporting bias – an updated review. *PloS One*, 8(7):e66844, 2013.

[44] Tyler J VanderWeele. On a square-root transformation of the odds ratio for a common outcome. *Epidemiology*, 28(6):e58–e60, 2017.

[45] Annie Franco, Neil Malhotra, and Gabor Simonovits. Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505, 2014.

[46] Jeffrey T Leek and Leah R Jager. Is most published research really false? *Annual Review of Statistics and Its Application*, 4:109–122, 2017.

[47] Daniele Fanelli. "Positive" results increase down the hierarchy of the sciences. *PLOS One*, 5(4):e10068, 2010.

[48] Denes Szucs and John PA Ioannidis. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, 15(3):e2000797, 2017.

[49] Jeff Leek. *tidypvals: This is a package with published p-values from the medical literature in tidied form.*, 2019. R package version 0.1.0.

[50] Abel Brodeur, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32, 2016.

[51] David Chavalarias, Joshua David Wallach, Alvin Ho Ting Li, and John PA Ioannidis. Evolution of reporting p values in the biomedical literature, 1990-2015. *JAMA*, 315(11):1141–1148, 2016.

[52] Tiago V Pereira and John PA Ioannidis. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of Clinical Epidemiology*, 64(10):1060–1069, 2011.

Supplement: Estimating Publication Bias in Meta-Analyses

Contents

1. Supplementary methods

1.1. Data extraction methods

We eliminated unpublished studies as follows. For many meta-analyses, the methods section suggested no attempts to include unpublished studies, in which case we simply included all point estimates in our analysis. When the methods section instead indicated some attempt to include unpublished studies, we contacted the meta-analysts or examined the reference lists of included studies in order to identify and exclude the unpublished studies. When possible, we treated as "published" any article published in a peer-reviewed journal or conference proceeding; however, we sometimes had to rely instead on the meta-analysts' own definitions of "published".

1.2. Dual coding and data entry quality checks

A team of six research assistants (Acknowledgments) and MBM extracted data, with two coders independently extracting data for every analyzed meta-analysis (except Metalab, which was singly coded because the publicly available datasets were already curated in an analysis-friendly format). The research assistants completed a customized 54-minute course of training videos (`https://osf.io/6dbhp/`) detailing how to extract data from the meta-analyses. To code journal tiers, we downloaded the full SciMago database of 2018 ratings (*Scimago Journal and Country Rank*, n.d.). We used an R script to standardize and merge journal titles from the SciMago database with those in our meta-analysis corpus. Some journals in our corpus did not have an exact match in the SciMago database because, for example, the title included a subtitle or section within the journal, the title we entered was abbreviated whereas the SciMago title was unabbreviated, the title had special characters or accents, the citation in paper was incomplete or misspelled, etc. For all such unmatched journals, we manually coded their Scimago ratings by splitting the work across four coders; finally, MBM manually checked and, if necessary, corrected every entry. We then used an R script to merge the resulting SJR dataset with our meta-analysis corpus, conducting sanity checks and data cleaning. For example, some journals had multiple, discrepant rankings because they had non-unique titles (e.g., *Surgery*); we removed these ambiguous journals from our SJR database so that they would result in missing data. Ultimately, 1.4% of point estimates had missing data on journal. Of the 1107 unique journals in our meta-analysis

corpus, rankings were hand-coded for 233 (23%) journals. Rankings were hand-coded for 11% of all point estimates in our corpus.

Upon the completion of data entry, we used an R script to check for extreme or incompatible values for all numerical entries, manually confirming or correcting each of these entries. We additionally compared the dual-coded datasets; where there were discrepancies, coders attempted to resolve these through discussion. When discrepancies of $\geq 5\%$ remained on analysis variables (e.g., the estimated selection ratio), MBM manually reviewed both coders' datasets, choosing one dataset for analysis by preferring datasets that: (1) proved to be correct on manual review of each point estimate and inference entry; (2) were prepped automatically in R by MBM rather than entered manually; and/or (3) exactly reproduced the paper's reported estimates when this was expected because the meta-analysis contained no unpublished studies. For two meta-analyses, limitations of analytic reproducibility precluded resolution of discrepancies (e.g., because there were inherent ambiguities in how to link study citations with study abbreviations in forest plots or because documentation regarding which studies were unpublished was unclear). Specifically, for PMID 26724178, the forest plot listed trial acronyms rather than unique publications, with each trial potentially yielding many separate publications. For PMID 28159391, most point estimates were from large public databases rather than publications, and point estimates from publications had no indication of which publication they were from. We excluded these meta-analyses from analysis as depicted in the PRISMA flowchart.

2. Supplementary results

2.1. Sensitivity analyses for violations of model assumptions

To assess for violations of the assumption that publication bias operates in favor of affirmative results (i.e., those with $p < 0.05$ and point estimates in the desired direction), we calculated and plotted one-tailed p-values from all studies in our dataset, treating the direction of the meta-analytic point estimate as the desired direction (Figure S1). The much larger mass of one-tailed p-values below 0.025 (50% of all p-values) versus those above 0.975 (5% of p-values) suggested that selection indeed was primarily one-directional, though a small mass above 0.975 suggests some weak two-tailed selection (i.e., selection favoring "significant" results regardless of sign). As a simple measure of apparent two-tailed selection in each meta-analyses, we calculated the ratio of the observed proportion of nonaffirmative studies

with one-tailed $p > 0.975$ to its expectation under the assumption that the p-values of all non-affirmative studies are uniformly distributed. Since nonaffirmative studies are those with a one-tailed $p > 0.025$, the expectation is therefore $0.025/0.975 \approx 0.0256$. In the below sensitivity analyses, we excluded meta-analyses for which this ratio exceeded 3.

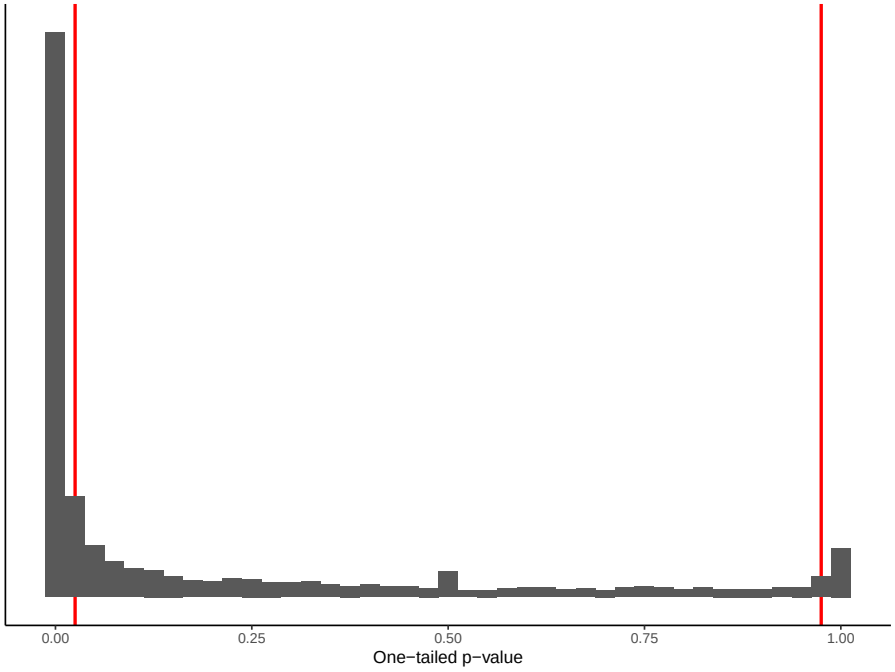

Figure S1: *One-tailed p-values from all meta-analyses, treating the direction of the meta-analytic point estimate as the desired direction. Red lines indicate the 0.025 and 0.975 thresholds, i.e., the thresholds at which the corresponding two-tailed p-value would be* < 0.05 *and in the desired direction and at which the two-tailed p-value would be* < 0.05 *but in the unanticipated direction.*

A second plausible threat to model assumptions is non-normal true effects, which we assessed by excluding meta-analyses for which a Shapiro test of the normalized point estimates yielded $p < 0.05$ (Hardy & Thompson, 1998; Shapiro & Wilk, 1965). This criterion is conservative in that the selection model assumes that the latent true effects are normal *prior to* selection due to publication bias, so meta-analyses with non-negligible publication bias may have normal true effects in the latent population despite having non-normal point estimates. A third potential threat is our inclusion of at least two meta-analyses (PMID 27416099 and 27835651) in which the authors coded as "0" the point estimate for any study that reported only a "nonsignificant" effect, creating point masses of estimates at exactly 0. These point masses would violate the normality assumption as well as produce a downward-biased estimate of the selection ratio. Our sensitivity analyses, below, excluded meta-analyses in which $> 5\%$

of estimates were coded as exactly 0. Finally, in one meta-analysis (PMID 28700728), the original dataset coded effects were coded in an internally inconsistent manner, rendering the direction of the point estimates meaningless. We additionally excluded this meta-analysis in sensitivity analyses. Table S1 summarizes the effects of applying each exclusion criterion, or all criteria simultaneously, on an overall estimate of the selection ratio. These results suggests that while many meta-analyses failed the stringent sensitivity analysis criteria, the resulting pooled point estimates were not substantially affected.

Possible threat	k	$\widehat{SR}$ [95% CI]	Max $\widehat{SR}$	q_{95}
Two-tailed selection	33	1.06 [0.79, 1.43]	7.80	2.47
Non-normality	29	1.42 [1.09, 1.86]	7.80	2.64
Point mass at zero	49	1.26 [0.97, 1.64]	54.37	4.15
Other	55	1.20 [0.94, 1.53]	54.37	3.72
All of above	15	1.43 [0.94, 2.17]	7.80	2.90

Table S1: *Effect of sensitivity analyses on overall estimate of selection ratio. k: number of meta-analyses included in the sensitivity analysis.*

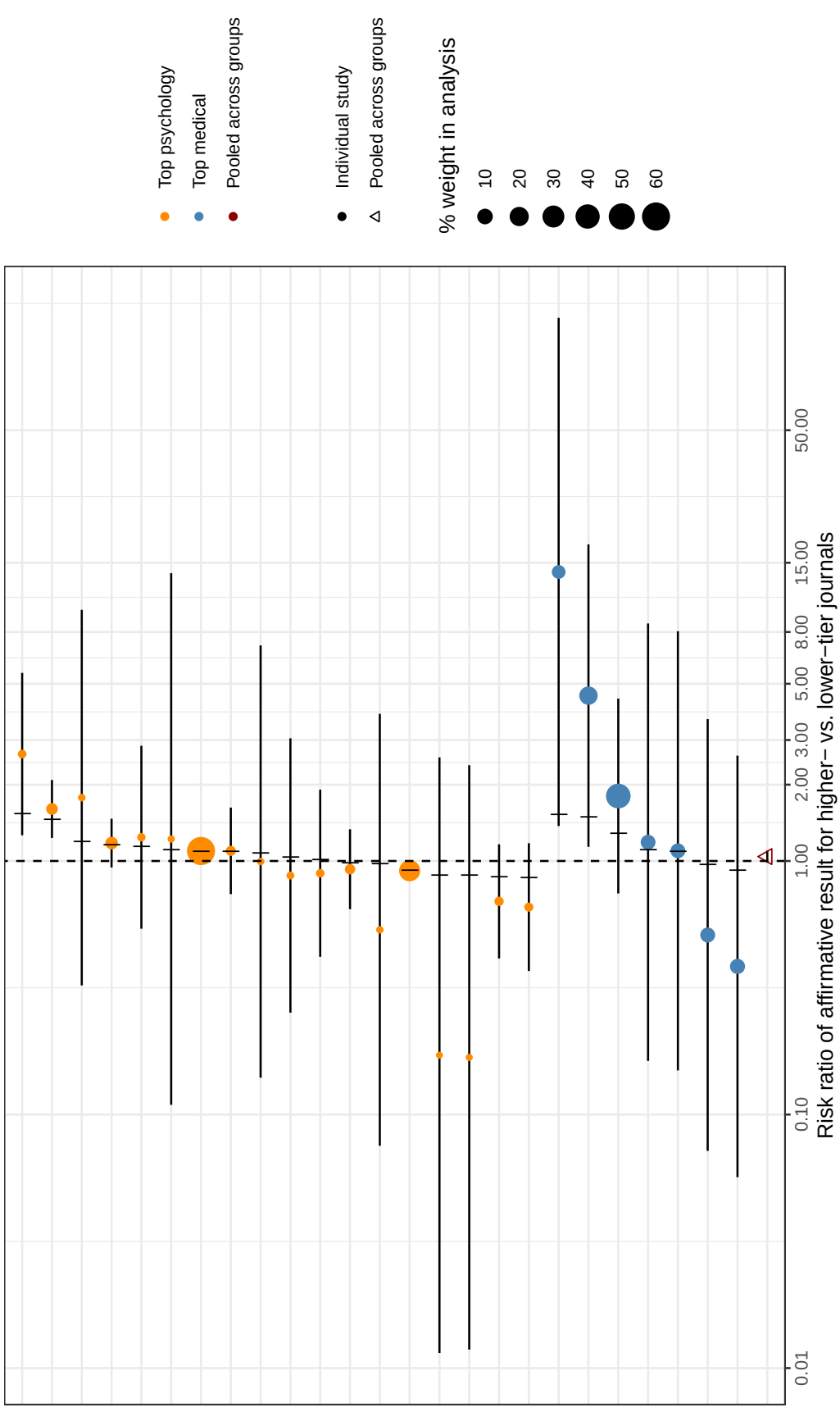

Figure S2: *Forest plot of risk ratios of an affirmative result comparing higher-tier to lower-tier journals. Within-study risk ratios were not estimable every meta-analysis because some meta-analyses had no higher-tier studies at all.*

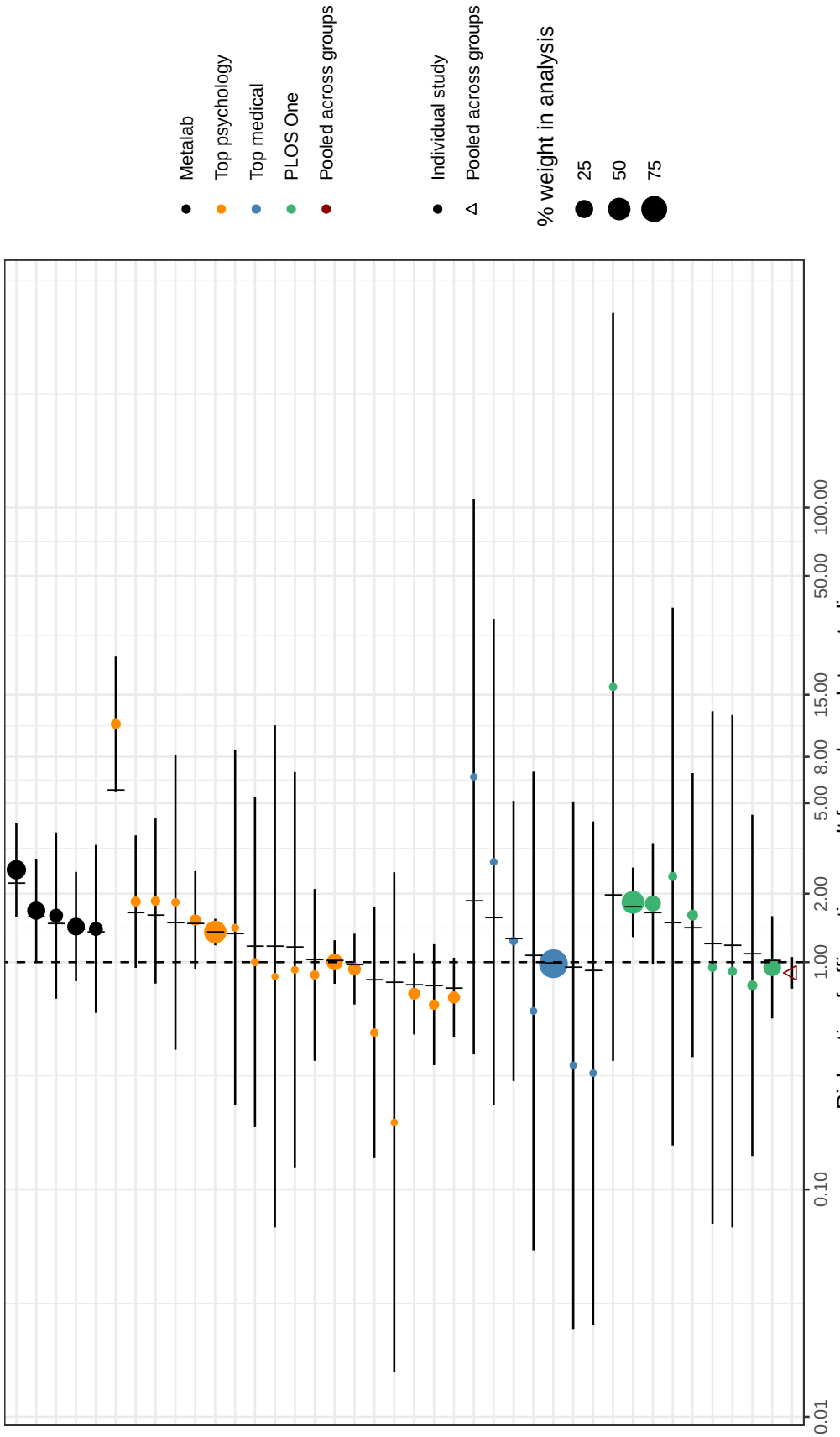

Figure S3: *Forest plot of risk ratios of an affirmative result comparing early to later studies.*

2.2. Sensitivity analyses excluding *Journal of Educational Psychology*

Because a single journal (*Journal of Educational Psychology*) contributed a particularly large percentage of higher-tier point estimates (47%), we conducted a sensitivity analysis in which we recoded this journal as "lower-tier". After doing so, we estimated that higher-tier results were 0.83 (95% CI: [0.76, 0.92]; $p = 3 \cdot 10^{-4}$) times as likely to be affirmative as lower-tier results.

3. Changes and additions to preregistered protocol

During article review, we decided to exclude network meta-analyses because these typically do not have study-level point estimates, though we made exceptions for network meta-analyses that also presented standard pairwise meta-analyses. We had originally planned to classify as "early" those studies that were "among the chronologically first three point estimates"; however, due to the large number of overlapping study years, this criterion appeared too lenient, so we adopted the criterion described in the main text. Regarding thresholds for "higher-tier" journals, we had initially planned to set the threshold for psychology to 3.25 and the threshold for medicine to 7.4 so that the lowest-ranked higher-tier journals in each category would be *Journal of Experimental Psychology: General* and *Annals of Internal Medicine*; we revised this threshold when new rankings became available after the preregistration was published. The preregistration indicated that we would consider the percentage of "statistically significant" results without specifying whether this would include results with point estimates in either direction or only affirmative results. For consistency with the selection models in main analyses, we chose to use affirmative status as the primary outcome and secondarily present analyses for "significant" results in either direction. The preregistration did not describe how we would conduct inference for the study-level measures, leading us to introduce the robust GEE models post hoc. All analyses described as sensitivity analyses or exploratory analyses were chosen post hoc.

References

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, *17*(8), 841–856.

Scimago journal and country rank. (n.d.). https://www.scimagojr.com/. (Accessed: 2019-07-08.)

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.