Check for updates

Theoretical/Methodological/Review Article

Clinical Psychological Science
1–15
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/21677026211049366
www.psychologicalscience.org/CPS

Effect Sizes Reported in Highly Cited Emotion Research Compared With Larger Studies and Meta-Analyses Addressing the Same Questions

Ioana A. Cristea[1,2], Raluca Georgescu[3], and John P. A. Ioannidis[2,4,5,6,7,8]

[1]Department of Brain and Behavioral Sciences, University of Pavia; [2]Meta-Research Innovation Center at Stanford (METRICS), Stanford University; [3]Department of Clinical Psychology and Psychotherapy, Babes-Bolyai University; [4]Department of Medicine, Stanford University; [5]Department of Epidemiology and Population Health, Stanford University; [6]Department of Biomedical Data Science, Stanford University; [7]Department of Statistics, Stanford University; and [8]Meta-Research Innovation Center Berlin (METRIC-B), QUEST Center for Transforming Biomedical Research, Berlin Institute of Health, Charité–Universitätsmedizin, Berlin, Germany

Abstract

We assessed whether the most highly cited studies in emotion research reported larger effect sizes compared with meta-analyses and the largest studies on the same question. We screened all reports with at least 1,000 citations and identified matching meta-analyses for 40 highly cited observational studies and 25 highly cited experimental studies. Highly cited observational studies had effects greater on average by 1.42-fold (95% confidence interval [CI] = [1.09, 1.87]) compared with meta-analyses and 1.99-fold (95% CI = [1.33, 2.99]) compared with largest studies on the same questions. Highly cited experimental studies had increases of 1.29-fold (95% CI = [1.01, 1.63]) compared with meta-analyses and 2.02-fold (95% CI = [1.60, 2.57]) compared with the largest studies. There was substantial between-topics heterogeneity, more prominently for observational studies. Highly cited studies often did not have the largest weight in meta-analyses (12 of 65 topics, 18%) but were frequently the earliest ones published on the topic (31 of 65 topics, 48%). Highly cited studies may offer, on average, exaggerated estimates of effects in both observational and experimental designs.

Keywords

emotion, highly cited, meta-analysis, metaresearch, citation bias, open data, preregistered

Received 5/5/21; Revision accepted 7/24/21

Highly cited (HC) studies are often considered to be the most valued and influential scholarship, which leads to an expectation that they should hopefully report the most accurate findings. However, meta-epidemiological investigations in some scientific fields have found that HC studies may report overestimated effects relative to larger or better designed studies (Ioannidis, 2005; Tajika et al., 2015) or to meta-analyses on the same topic (Ioannidis & Panagiotou, 2011). In addition, influential studies often produce substantially larger or contradictory effects relative to subsequent preregistered replication attempts (Camerer et al., 2018; Klein et al., 2018; Open Science Collaboration, 2015; Wagenmakers et al., 2016).

Multiples sources of bias may contribute to effect size inflation (Fanelli et al., 2017; Ioannidis et al., 2008). A major concern is that when research findings are incentivized to pass a prescribed threshold of statistical significance to be published (*publication bias*) and

Corresponding Author:
Ioana A. Cristea, Department of Brain and Behavioral Sciences, University of Pavia
Email: ioana.cristea@unipv.it

research designs have suboptimal statistical power, published effect sizes are inflated on average (Bakker et al., 2012; Button et al., 2013; Gelman, 2018; Ioannidis, 2008). In addition, flexibility in analytical choices (Simmons et al., 2011) can lead to a large "vibration of effects" (i.e., the range of possible effects obtained for different analysis specifications estimating the same association) that, when combined with selective reporting, can lead to an upward bias for published effects (Patel et al., 2015; Steegen et al., 2016). Finally, influential stakeholders within the scientific ecosystem, such as funders and journals, exert a preference for aesthetically appealing ("positive," "clean," or "novel") results (Nosek et al., 2012), which could lead to preferential citation (*citation bias*) of studies that report larger effects compared with those reporting smaller or null effects (Cristea & Naudet, 2018; Gøtzsche, 1987; Greenberg, 2009).

Although systematic investigations of effect-size inflation in HC articles in the social and behavioral sciences are lacking, indirect evidence from replication studies suggests that effects reported by HC studies may be exaggerated. For example, three large-scale studies have found that effects reported in multilaboratory preregistered replication attempts are on average 49% to 66% smaller than corresponding effects reported in previously published research (Camerer et al., 2016, 2018; Open Science Collaboration, 2015). Many, but not all, of the included original studies were HC, and they were all published in high-profile journals. Other multilaboratory replication efforts specifically targeting influential psychology studies often report smaller (and sometimes null) effects relative to original studies (Klein et al., 2014, 2018; Wagenmakers et al., 2016). These replication efforts did not systematically target specifically the most HC articles—even though some of the assessed work was HC. Moreover, they have also focused predominantly on randomized experiments. However, there are many other studies that attract a lot of attention and citations, including diverse observational associations, biomarkers or predictive markers, and more. It would be important to assess whether HC studies covering such a broad spectrum of designs have inflated effect sizes and, if so, the size of the inflation compared with other studies on the same questions that do not get so many citations.

Hence, the goal of the present study was to investigate whether effect sizes reported in HC emotion research are greater relative to larger studies and meta-analyses addressing the same questions. We focused on emotion research because it is a major topic domain in psychology with a breadth of content and research designs and covers both highly exploratory basic research and applied research with clinical implications. Our goal was to gauge the extent to which effects differed between HC studies and summary effects from meta-analyses and the larger studies on the same topic. We also wanted to map the timing of publication of HC studies, largest studies, and other studies on the same topic.

Method

We adopted the approach of previous similar investigations in clinical research (Ioannidis, 2005; Ioannidis & Panagiotou, 2011) and psychiatry (Tajika et al., 2015). Changes to the preregistered study protocol are detailed in the Supplemental Material available online.

Identification and selection of target HC articles

The database Scopus was searched through October 8, 2019, using keywords generically related to "emotion," "mood," "anxiety," or "depression" present in the title, abstract, or keywords.

Eligible records reported on primary data that could be used for generating effect sizes in human participants, mentioned findings related to emotions in the abstract (even if these were peripheral to the goals of the study), had an experimental (i.e., randomized) or observational design, and had been cited at least 1,000 times in Scopus as of the date of the search. Articles in which the abstract made no mention of emotion or focused exclusively on biomedical, molecular, or other aspects not related to emotional disorders or conditions were excluded. However, articles that were found to mention emotion during abstract inspection, even if in a peripheral role (e.g., as one of many secondary outcomes, a component in a model), were included.

We also excluded (a) meta-analyses and other articles using secondary data, (b) observational studies focused on prevalence, (c) studies describing the development or subsequent validation of scales, and (d) estimations of disease burden, such as the Global Burden of Disease.

One researcher (I. A. Cristea) screened all records with at least 1,000 citations by title and abstract and selected those that mentioned emotion and described observational (including pre/post designs and nonrandomized studies of various associations) and experimental (including all studies in which participants were randomized to an intervention or to different modalities of an independent variable) designs.

Identification and selection of meta-analyses

For each eligible observational or experimental HC record, we searched for the most recent meta-analysis

including effect-size data from any finding in the article, provided it was related with emotion. In cases in which the HC article mentioned emotion in a peripheral role, an eligible meta-analysis had to report effect size related to the emotion finding and not to the article's other findings.

Meta-analyses for each target article were identified by downloading the most recent 2,000 records citing the target study in the form of a searchable .csv file. We then used the "find" command in a document processor to search for the text string "meta-analy*" in the title, author, or index keywords. Citing records were screened starting with the most recent ones and moving downward on the list. Whenever a potentially eligible meta-analysis was identified, the full text was retrieved and manually searched to identify whether (a) the HC study was included and (b) an effect size of interest from the HC study was reported. If these criteria were not satisfied, we moved down the list of citing articles chronologically until identifying another eligible meta-analysis. Meta-analyses that substituted the HC study with a larger study that encompassed it or with another publication on the same sample were eligible. In these cases, we planned to recalculate the effect size from the original report, if possible.

For eligible records that described more than one meta-analysis (i.e., reported more than one forest plot) including the target HC study, we chose the one with the highest number of studies or, if there were ties in this regard, the one that appeared first in the text, provided it reported a finding related to emotion.

One researcher (I. A. Cristea) searched citing records and identified meta-analyses.

Data extraction

For each matching meta-analysis, we coded information about publication year, meta-analysis model used (fixed or random), total number of included effect sizes in the selected forest plot, effect-size measure (e.g., mean difference, standardized mean difference [SMD], correlation, odds ratio [OR], risk ratio [RR], hazard ratio [HR]), earliest study (by publication year) in the forest plot, effect sizes and 95% confidence intervals (CIs) for the HC and largest study, and summary effect sizes and 95% CIs in the meta-analysis. When the meta-analysis reported different models of estimating effect size, we preferred random effects. The largest study was defined as the study with the lowest standard error in the matching meta-analysis. To select the largest study, we relied on the following succession of information, if reported: (a) weights in the forest plots, followed by (b) standard errors/variance associated to individual effect sizes, followed by (c) recalculation of the 95% CI width for those individual effect sizes in which the CI appeared visually smaller in the forest plot, and finally, (d) study sample size. If more studies with the same weight or standard error were included, sample size was used to break the tie.

If more studies, including the HC study, were the earliest in the forest plot (published in the same year), the HC study was considered the earliest.

When the forest plot included only graphic information, we attempted to contact the authors or used tools such as WebPlotDigitizer (https://automeris.io/WebPlotDigitizer/) to reconstruct the data from the plots.

Outcomes

All outcomes were assessed separately for observational and experimental designs.

The primary outcome was the degree of agreement between (a) the effect size of the HC study and the summary effect size of the matching meta-analysis and (b) the effect size of the HC study and the effect size in the largest study in the matching meta-analysis. To this purpose, we calculated the ratios of odds ratios (RORs), as detailed in the Data Analysis section.

This outcome is reported both nominally, as the percentage of topics in which the 95% CI of ROR included 1, and statistically, as the meta-analytical aggregate across topics, separately for experimental and observational studies.

Secondary outcomes were the percentages of HC studies with effect sizes that differed by 2-fold (ROR $\geq$ 2 or $\leq$ 0.5) or 4-fold (ROR $\geq$ 4 or $\leq$ 0.25) from the effect size in the matching meta-analysis and respective largest study in the meta-analysis.

Data analysis

Analyses were performed in Microsoft Excel for Mac (Version 16.43) and STATA/SE for Mac (Version 16.1; programs *admetan* and *metaeff*). Scatterplots were constructed in the R software environment (R Core Team, 2020) using RStudio (Version 1.2.5033; RStudio Team, 2019) and the *lessR* package (Version 3.9.8; Gerbing, 2020).

For each identified meta-analysis, we extracted the effect size and standard error or 95% CI reported for the HC study, the summary effect size, and the effect size of the largest study in the meta-analysis. The preferred meta-analytic estimate was the OR. Effect sizes were extracted as reported in the meta-analysis without retrieving the primary studies. When meta-analyses reported estimates other than the OR, we employed standard procedures for converting estimates into ORs. SMDs, including Hedges's *g*, were transformed to

naturally logarithmic ORs using the Chinn transformation (Chinn, 2000). Correlation coefficients were first converted into SMDs (Polanin & Snilstveit, 2016) and then into ORs. For RRs that could not be converted into ORs without estimates of baseline risk, often not reported, we first checked whether study-level event data (e.g., a 2 × 2 table) were reported. If yes, we extracted them and reran the meta-analysis with effect sizes expressed as ORs using the authors' specified meta-analytic model. If neither baseline risk nor event data were provided, remaining RRs were treated as ORs in the main analyses and excluded in sensitivity analyses. Likewise, HRs were assimilated to ORs. For continuous outcomes expressed as mean differences and standard errors or CIs, we also reran the meta-analysis to produce SMDs.

For meta-analyses that reported data by subgroups, we took the pooled estimate (i.e., across subgroups) if available and the estimate in the largest subgroup including the HC study if the pooled estimate across all subgroups was not reported. For forest plots that included separate effect sizes from the HC or largest study (e.g., different subgroups or outcomes), we first pooled these distinct estimates under a fixed-effects model and used that estimate for further analyses, whereas the summary estimate remained the one reported in the meta-analysis.

To assess the magnitude of the differences for each pair (HC vs. summary estimate; HC vs. largest study), we computed the RORs using the Altman-Bland approach (Altman & Bland, 2003). In brief, RORs were obtained by dividing the OR of the HC article by the (a) summary effect size of the meta-analysis and (b) the effect size in the largest study.

To ensure coherence across studies, effect sizes were coined (i.e., the sign was inverted) when necessary. For experiments, coining was performed so that an ROR greater than 1 implied that the intervention or experimental manipulation had more favorable results than control. For observational studies, exposures were coined to represent values over 1 for the HC study so that an ROR greater than 1 meant that the effect size in the HC study was larger than the one in the meta-analysis or larger study. For each comparison of the HC study and meta-analysis and HC study and largest study, we noted whether RORs were statistically significant (i.e., the 95% CI did not include 1) and whether estimates from the HC study differed by at least 2-fold (ROR ≥ 2 or ≤ 0.5), at least 4-fold (ROR ≥ 4 or ≤ 0.25), or more.

We also conducted meta-analyses of RORs separately for experimental and observational designs. Although in the protocol we planned both fixed- and random-effects models for meta-analyses of RORs, given the substantial clinical heterogeneity, we reported only a random-effects model. We used a random-effects model with the Paule and Mandel estimator (Paule & Mandel, 1989), recommended for dichotomous outcomes in the presence of high heterogeneity (Veroniki et al., 2016). Although not specified in the protocol, for comparisons with the largest study, cases in which the largest study coincided with the HC study were excluded. Heterogeneity was assessed with the between-topics variance τ^2, I^2, and its 95% CI estimated using the Q-profile method (Viechtbauer, 2007). Because some clinical psychologists may be accustomed to SMD rather than OR metrics for expressing effects, we also transformed summary RORs from the main analysis in differences of SMDs (dSMDs) by applying the conversion formula described by Chinn (2000) using the natural logarithm (ln) of the ROR. For the ROR of the HC study versus the summary estimate of the meta-analysis (MA in the equation), we have:

$$\ln\left(\mathrm{ROR}_{\mathrm{HC\ MA}}\right) = \ln\left(\frac{\mathrm{OR}_{\mathrm{HC}}}{\mathrm{OR}_{\mathrm{MA}}}\right) = \ln\left(\mathrm{OR}_{\mathrm{HC}}\right) -$$

$$\ln\left(\mathrm{OR}_{\mathrm{MA}}\right) = 1.81 \times \mathrm{SMD}_{\mathrm{HC}} - 1.81 \times \mathrm{SMD}_{\mathrm{MA}} = 1.81\,\mathrm{dSMD}.$$

Therefore, $\mathrm{dSMD} = \ln(\mathrm{ROR}_{\mathrm{HC\ MA}}) \div 1.81$.

The standard errors can be computed by the same formula, $SE(\mathrm{dSMD}) = SE(\mathrm{ROR}_{\mathrm{HC\ MA}}) \div 1.81$. We also reported estimates as dSMDs for the topics in which the SMD was the effect measure used in the selected meta-analysis.

Sensitivity analyses were performed by repeating the main analyses (a) excluding studies for which HRs and RRs were considered to be ORs, (b) limited to the topics in which the HC study was the earliest published on the topic, (c) limited to the topics in which the largest study was published later than the HC study, and (d) restricted to HC studies in which the abstract mentioned the outcome and exposure/intervention extracted from the matching meta-analysis. This last analysis was not preregistered and was added post hoc to verify whether the finding selected for evaluation was considered central in the HC study. Finally, because we observed extremely large heterogeneity for observational designs, we added a series of nonpreregistered exploratory sensitivity analyses for this cohort (e) limited to topics in which both the exposure and outcome are clinical manifestations (e.g., depression, anxiety, insomnia), demographic variables (e.g., gender), or major life events (e.g., adverse events, childhood abuse); (f) limited to topics in which either exposure or outcome are nonclinical or surrogate measurements (e.g., genes, neuroimaging, cognitive tasks); and (g) limited to topics in which the matching meta-analysis was lower variance or higher variance compared with the median of the entire sample. For this analysis,

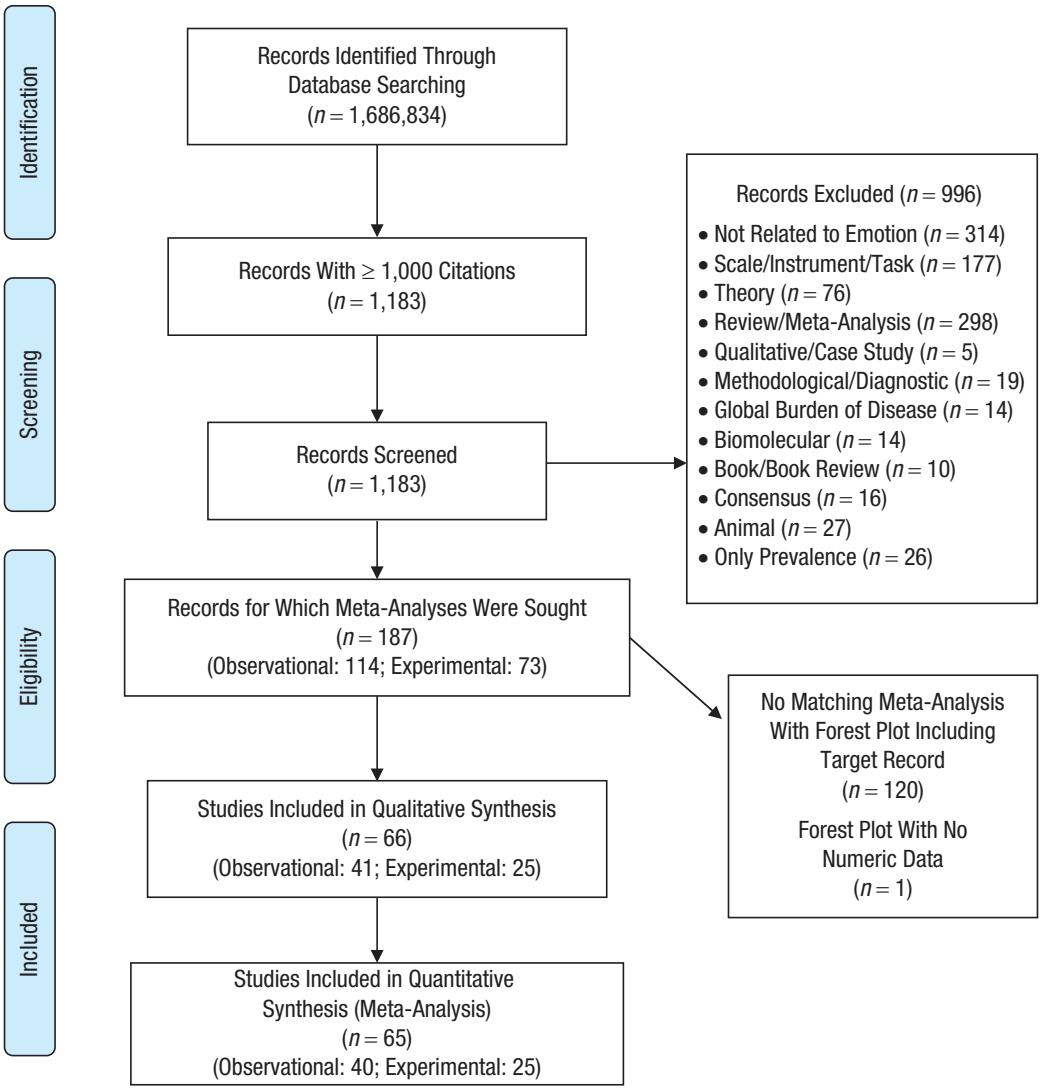

Fig. 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram of the study-selection process.

we calculated the median standard error of the log OR for the entire cohort of meta-analyses and used the median to dichotomize the sample into topics in which the standard error of the log OR in the matching meta-analysis was below and above the median.

Results

Selection of target HC articles and matching meta-analyses

The search produced 1,686,834 records, of which 1,183 had at least 1,000 citations. From these, 187 studies were selected (114 observational and 73 experimental; for the Preferred Reporting Items for Systematic Reviews and Meta-Analyses [PRISMA] flow diagram, see Fig. 1). Twenty-seven studies (14%) had more than 2,000 citations, and as per protocol and owing to Scopus limitations on download of citing records, we screened through only the first 2,000 most recent citations until identifying a matching meta-analysis. This procedure failed to identify a matching meta-analysis for 19 of these 27.

We contacted authors of three meta-analyses in which forest plots did not contain effect-size data or were incomplete (i.e., presented only a subgroup) and retrieved data for two meta-analyses. The remaining meta-analysis (McKinnon et al., 2009) was excluded. Therefore, we identified matching meta-analyses with study-level effect-size data for 41 of 114 (36%) observational studies (37 unique meta-analyses, four of which contained more than one HC study) and 25 of 73 (34%)

experimental studies (22 unique meta-analyses, three of which contained two different HC studies).

Characteristics of the sample

The 41 observational studies were published between and 1972 and 2013, and citation counts ranged from 1,001 to 5,497 (*Mdn* = 1,357, interquartile range [IQR] = 1,087–1,769). The 25 experimental studies spanned 1989 to 2006, and citation counts ranged from 1,126 to 2,374 (*Mdn* = 1,426, IQR = 1,290–1,723). Matching meta-analyses were published between 1998 and 2019 for observational studies and between 2014 and 2019 for experimental studies. Twenty-eight of 37 meta-analyses for observational studies (76%) and 21 of 22 (95%) experimental studies were published after 2015 (see Tables S1 and S2 in the Supplemental Material).

For observational studies, 17 meta-analyses used ORs, 14 used SMDs (five used Hedges's *g*), seven used RRs, one used HR, one used mean difference, and one used standardized coefficients from linear regression. In this last case, we did not have enough information to convert or recalculate the regression coefficient, and the meta-analysis (Martinez-Calderon et al., 2019) was excluded, which left a total of 40 meta-analyses for quantitative synthesis.

There were a few special cases. In the case of one meta-analysis involving individual patient data (Culverhouse et al., 2018), the HC study was eligible for inclusion but did not provide primary data. As per protocol, we used the estimates from the meta-analysis for the summary and the largest study estimates and recalculated the effect for the HC study from the primary report, using coining so it would represent the same contrast as the meta-analysis. For another HC study (Regier et al., 1990), the meta-analysis (Lai et al., 2015) included data from a pooled analysis (Swendsen et al., 1998) combining the cohort reported in the HC study with other cohorts. The HC study did not report the sufficient data for effect-size calculation, but the meta-analysis included separate estimates for the HC study cohort (Epidemiologic Catchment Area), which we extracted. Another meta-analysis (Reising et al., 2019) included a larger study that contained the HC study (Odgers et al., 2008). The HC study did not report sufficient data for effect-size calculation. We substituted its estimate with the one from the overlapping larger study (Odgers et al., 2008) reported in the meta-analysis. Because the original HC study reported only on males and the overlapping study included separate estimates for males and females, we used only the former. We conducted additional sensitivity analyses excluding all these special cases.

For experimental studies, 13 meta-analyses reported SMDs (six reported Hedges's *g*), four reported ORs, four reported RRs, three reported Pearson correlation coefficients (*r*), and one reported HR. For one HC study describing the Enhancing Recovery in Coronary Heart Disease Patients randomized trial (Writing Committee for the ENRICHD Investigators, 2003), the corresponding meta-analysis (Richards et al., 2017) combined data from all trial publications. As per protocol, we used summary and largest study estimates from the meta-analysis and recalculated the HC study effects using the primary report.

Primary outcomes and meta-analysis

Observational studies. Effect estimates were recalculated (n = 2) or converted (n = 15) for 16 meta-analyses. For six meta-analyses, RRs and HRs were assimilated to ORs in computing RORs. Effect estimates were coined for eight meta-analyses. For 27 of 40 (67.5%) HC studies, we rated the abstract as describing the finding extracted from the matching meta-analysis. Twenty-five of 40 (62.5%) HC studies were the earliest or conducted within 3 years of the earliest study in the meta-analysis (see Tables 1 and 2 and Fig. 2).

In 27 of 40 HC studies (67.5%), estimates were nominally larger (i.e., ROR > 1) than the summary effect in the corresponding meta-analysis (Fig. 2). In 12 of 40 HC studies (30%), effects were statistically significantly different from the summary estimate, and in 10 cases, RORs were greater than 1. The difference was at least 2-fold for 15 (37.5%) pairs and at least 4-fold for six (15%) pairs. The summary ROR (see Fig. S1 in the Supplemental Material) across all topics was 1.42 (95% CI = [1.09, 1.87]) with extremely high heterogeneity (τ^2 = 0.55, I^2 = 98%, 95% CI = [95%, 99%]). This was equivalent to a dSMD of 0.19 (95% CI = [0.04, 0.34]; Table 2). RORs were somewhat larger for topics in which the HC was the earliest study (n = 21; ROR = 1.77, 95% CI = [1.07, 2.94], τ^2 = 1.07) and those in which the HC study abstract mentioned the exposure and outcome used in the meta-analysis (n = 27; ROR = 1.60, 95% CI = [1.08, 2.36], τ^2 = 0.77), but heterogeneity remained very large. For topics in which the matched meta-analysis reported effects as SMDs (n = 15), estimates were higher (dSMD = 0.45, 95% CI = [0.04, 0.86]) and had extremely high heterogeneity (τ^2 = 1.66).

Heterogeneity was significantly reduced in exploratory sensitivity analyses of topics in which both exposure and outcome were clinical manifestations, demographic variables, or major life events (n = 25; ROR = 1.16, 95% CI = [0.94, 1.42], τ^2 = 0.17). Heterogeneity was also contained in analyses (n = 20) circumscribed to the meta-analyses with lower variance (i.e., under the median variance of the entire cohort of meta-analyses; ROR = 1.18, 95% CI = [0.92, 1.52], τ^2 = 0.26).

Table 1. Meta-Analytic Estimates and Sensitivity Analyses of Ratio of Odds Ratios for Observational Designs

Comparison	*N*	ROR	95% CI	I^2 [95% CI]	τ^2
HC study compared with summary estimate	40	1.42	[1.09, 1.87]	98 [95, 99]	0.55
Excluding HC study estimates recalculated from primary articles[a]	37	1.47	[1.08, 1.99]	98 [95, 99]	0.64
RRs and HRs substituted for ORs[b]	34	1.54	[1.11, 2.14]	98 [95, 99]	0.69
Excluding coined estimates	32	1.30	[0.97, 1.74]	97 [92, 99]	0.51
HC study is the earliest study	21	1.77	[1.07, 2.94]	98 [95, 99]	1.07
Meta-analytic comparison mentioned in HC study abstract	27	1.60	[1.08, 2.36]	95 [86, 98]	0.77
Exposure and outcome clinical/demographic/ life events	25	1.16	[0.94, 1.42]	91 [80, 97]	0.17
Exposure or outcome nonclinical/ surrogate	15	2.22	[1.11, 4.45]	99 [97, 100]	1.55
Variance of meta-analysis < median variance of the cohort	20	1.18	[0.92, 1.52]	98 [92, 99]	0.26
Variance of meta-analysis > median variance of the cohort	20	1.91	[1.14, 3.21]	87 [66, 95]	0.93
HC study compared with largest study	35	1.99	[1.33, 2.99]	98 [97, 99]	1.17
Excluding HC study estimates recalculated from primary papers[a]	32	2.08	[1.33, 3.28]	97 [94, 99]	1.33
RRs and HRs substituted for ORs[b]	30	2.27	[1.41, 3.66]	97 [93, 98]	1.38
Excluding coined estimates	28	1.62	[1.06, 2.48]	98 [96, 99]	1.04
HC study predates largest study	33	1.99	[1.29, 3.07]	98 [96, 99]	1.25
Meta-analytic comparison mentioned in HC study abstract	26	2.14	[1.31, 3.48]	97 [93, 99]	1.22
Exposure and outcome clinical/demographic/ life events	22	1.49	[0.97, 2.30]	98 [96, 99]	0.81
Exposure or outcome nonclinical/ surrogate	13	3.31	[1.50, 7.32]	89 [71, 97]	1.69
Variance of meta-analysis < median variance of the cohort	17	1.71	[1.22, 2.41]	98 [92, 99]	0.40
Variance of meta-analysis > median variance of the cohort	18	2.51	[1.17, 5.41]	91 [82, 97]	2.12

Note: ROR = ratio of odds ratio; CI = confidence interval; HC = highly cited; HR = hazard ratio; RR = risk ratio; OR = odds ratio.
[a]Excluded HC studies: Caspi (2003), Regier (1990), and Moffitt (2002). [b]The comparison excludes cases in which RRs and HRs could not be converted and were considered equivalent to ORs.

Five HC studies (12.5%) were also the largest, which left 35 for further ROR analyses. In 29 of 35 (83%) cases, HC study estimates were greater than those in the largest study in the meta-analysis (Fig. 2). In 17 cases, RORs comparing estimates were statistically significantly different from 1 (49%); in 13 of these cases, ROR was greater than 1. RORs were at least 2-fold for 17 of 35 cases (49%) and at least 4-fold in 10 of 35 (29%) cases. The summary ROR (see Fig. S2 in the Supplemental Material) was 1.99 (95% CI = [1.33, 2.99]) and had extremely high heterogeneity ($\tau^2 = 1.17$, $I^2 = 98\%$, 95% CI = [97%, 99%]). This corresponded to a dSMD of 0.38 (95% CI = [0.15, 0.61]; Table 2). The summary ROR was similar in sensitivity analyses restricted to topics in which the HC study predated the larger study ($n = 33$; ROR = 1.99, 95% CI = [1.29, 3.07]) and had similarly high heterogeneity ($\tau^2 = 1.25$). In exploratory sensitivity analyses (Table 1) on topics in which both exposure and outcome were clinical, summary RORs were reduced, and heterogeneity was more contained (ROR = 1.49, 95% CI = [0.97, 2.3], $\tau^2 = 0.81$). Heterogeneity was substantially reduced in analyses limited to meta-analyses with lower variance (ROR = 1.71, 95% CI = [1.22, 2.41], $\tau^2 = 0.40$).

Experimental studies. Effect estimates were recalculated ($n = 4$) or converted ($n = 16$) for 20 meta-analyses, and for one meta-analysis, the HR was considered equivalent to the OR. Effect estimates were coined for eight meta-analyses. Fifteen of 25 (60%) HC studies were the earliest or conducted within 3 years of the earliest study in the meta-analysis. The abstract of 21 of 25 (84%) HC

Table 2. Meta-Analytic Primary Analyses Estimates and Sensitivity Analyses Expressed as Differences in Standardized Mean Differences

Design and comparison	*N*	dSMD	95% CI	τ^2
Observational				
HC study vs. summary estimate	40	0.19	[0.04, 0.34]	0.55
Meta-analyses reporting SMDs	15	0.45	[0.04, 0.86]	1.66
HC study vs. largest study	35	0.38	[0.15, 0.61]	1.17
Meta-analyses reporting SMDs	14	0.73	[0.28, 1.18]	1.91
Experimental				
HC study vs. summary estimate	25	0.14	[0.007, 0.27]	0.20
Meta-analyses reporting SMDs	13	0.24	[0.05, 0.44]	0.25
HC study vs. largest study	18	0.39	[0.26, 0.52]	0.04
Meta-analyses reporting SMDs	9	0.41	[0.17, 0.65]	0.18

Note: dSMD = difference in standardized mean differences; CI = confidence interval; HC = highly cited; SMD = standardized mean difference.

studies described the intervention and outcome used in the meta-analysis (see Tables 2 and 3 and Fig. 3).

For 17 of 25 (68%) HC studies, estimates were nominally larger than summary estimates of matching meta-analyses (Fig. 3). The ROR of the HC study compared with the summary estimate was statistically significantly different from 1 in six of 25 cases (24%), three of which had RORs greater than 1. The estimates from the HC study differed by at least 2-fold (i.e., ROR ≥ 2 or ≤ 0.5) in five cases (20%) and by at least 4-fold in one case (4%). The summary ROR (see Fig. S3 in the Supplemental Material) was 1.25 (95% CI = [0.97, 1.61]) and had substantial heterogeneity ($\tau^2 = 0.25$, $I^2 = 73\%$, 95% CI = [53%, 87%]). The ROR corresponded to a dSMD of 0.14 (95% CI = [0.007, 0.27]). For topics in which the matched meta-analyses reported effects as SMDs, summary estimates were higher (dSMD = 0.24, 95% CI = [0.05, 0.44]). Sensitivity analyses limited to topics in which the HC study was the earliest study ($n = 10$) resulted in a similar summary ROR of 1.33 (95% CI = [1.08, 1.64]) and had no between-topics heterogeneity ($\tau^2 = 0$). Analyses of topics in which the HC study abstract mentioned the intervention and outcome extracted from the matching meta-analysis ($n = 21$) led to a similar summary ROR of 1.30 (95% CI = [0.97, 1.73], $\tau^2 = 0.33$).

Seven HC studies (28%) were also the largest in the matching meta-analysis, which left 18 studies for ROR analyses. Estimates from the HC study were nominally higher than those from the largest study for all 18 studies (Fig. 3), and for six of 18 (33%) studies, RORs were statistically significantly different from 1. RORs were at least 2-fold in seven of 18 (39%) cases and at least 4-fold in one of 18 (6%) cases. The summary ROR (see Fig. S4 in the Supplemental Material) of the HC study compared with the largest study was 1.81 (95% CI = [1.39, 2.36]) and had moderate heterogeneity ($\tau^2 = 0.09$, $I^2 = 29\%$, 95% CI = [0%, 68%]). One HC study postdated the largest study. Analyses restricted to the cases in which the HC study ($n = 17$) predated the largest study yielded a larger summary ROR of 1.85 (95% CI = [1.39, 2.46], $\tau^2 = 0.10$).

Discussion

Reports that collect extreme numbers of citations can be very influential in shaping the scientific literature and often also inform crucial decisions about which research to conduct, publish, or finance. Hence, the validity of their claims is paramount. Although there is no perfect means to evaluate validity, placing the results of HC studies against those of meta-analyses and of the largest studies on the same topic can offer valuable comparative insights. In a large, field-wide survey of emotion research, we showed that HC studies report more prominent effects compared with meta-analyses and larger studies on the same topic. For observational designs, HC studies produced effects about 1.4-fold higher on average than those from meta-analyses and almost 2-fold higher than those from the largest studies. For experimental designs, the average difference was around 1.3-fold for summary estimates and 2-fold in comparisons with the largest study. Translated in dSMDs, an estimate more habitually used by clinical psychologists, HC observational studies produced effects higher by 0.19 compared with summary meta-analytic estimates and by 0.38 compared with the largest study. The differences were similar for experimental designs (0.14 compared with the summary estimate and 0.39 compared with the largest study).

These average differences need to be viewed with great caution because there was extremely prominent heterogeneity across topics. Heterogeneity was extremely high for observational designs and more moderate for experimental studies. Heterogeneity was considerably reduced (a) in exploratory sensitivity

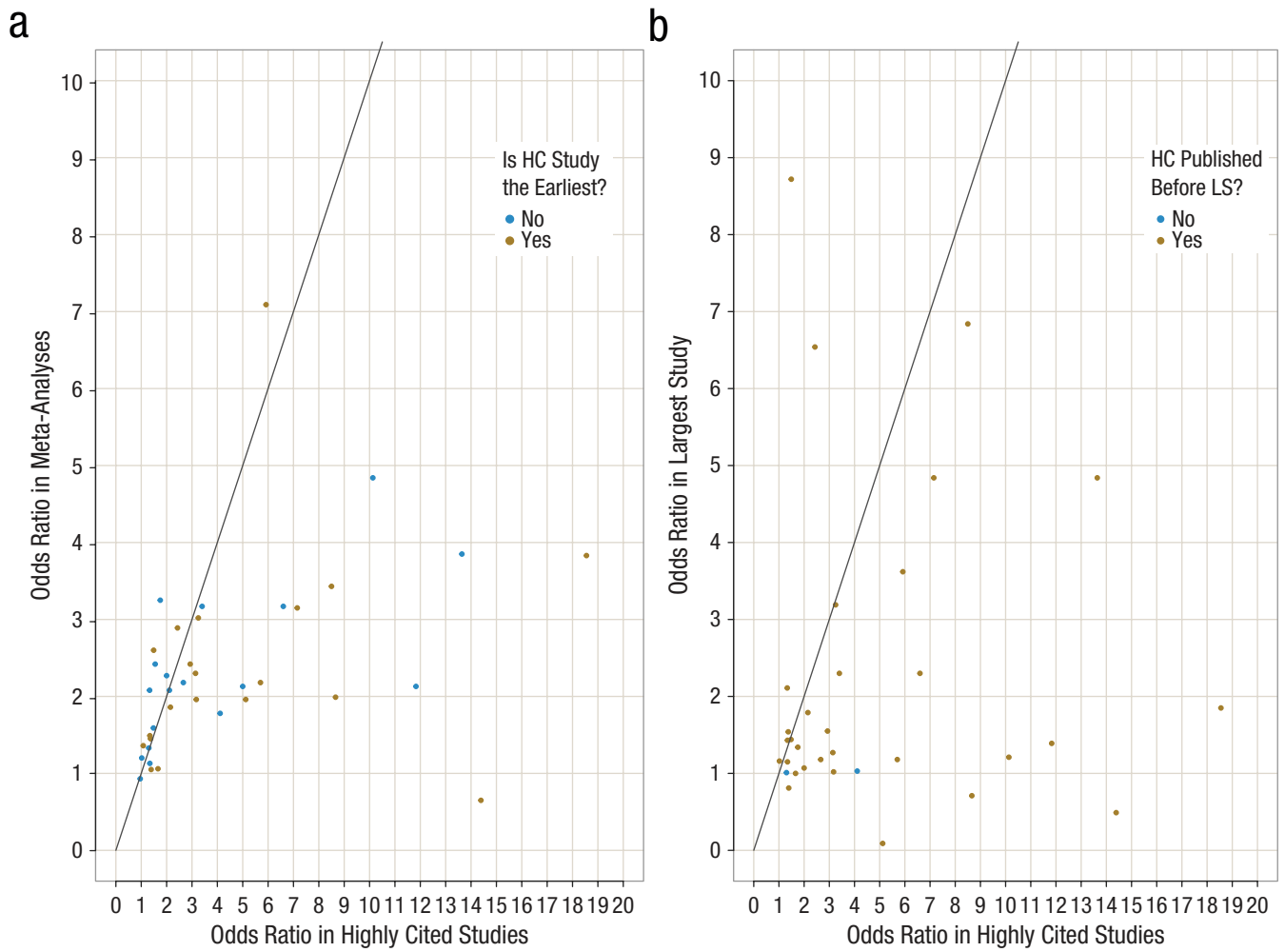

Fig. 2. Scatterplots showing the relation between odds ratios in the highly cited studies and (a) the corresponding summary estimates in the meta-analyses and (b) the corresponding largest-study estimates in observational designs. The diagonal lines show where the points would fall if the effects were equal. Not shown are two very large outliers (odds ratios of 883 and 49 in highly cited studies). For (b), five topics in which the highly cited study was the largest study are also not shown. LS = largest study.

analyses that were limited to topics in which both the exposure and outcome were clinical or demographic—or involved major life events—and (b) when considering the meta-analyses with the lower variance (i.e., below the median variance of the entire cohort of matching meta-analyses). In these analyses, differences between estimates were reduced to around 1.2-fold compared with meta-analyses and to 1.5 to 1.7 compared with larger studies and were, in most cases, nonsignificant (the CI around RORs included 1, albeit narrowly). Therefore, although HC studies may be expected to report larger effects on average, it is not possible to predict in advance for which topics this will be most pronounced and for which topics HC studies may not have larger effects at all. It is impossible to "correct" the effect estimates of an HC study by using some standard inflation factor.

We also examined the timing of publication of the HC reports compared with the other studies and with the largest studies on the same topic. HC studies are sometimes the first ones on the topic, and thus they would be the earliest published among the studies included in a meta-analysis. This pattern occurred in almost half (31 of 65, 48%) of the topics that we examined. However, in approximately 40% of cases, HC studies were published later or even substantially later (i.e., > 3 years after). Their high citation profile may reflect early publication ("being the first"), some citation bias favoring extreme results, or a combination thereof. Relatedly, the HC study predated the largest study in about two thirds of the pairs for observational designs and in all but one for experimental ones. Sensitivity analyses limited to topics in which the HC studies were the first ones mirrored the main analyses. If anything,

Table 3. Meta-Analytic Estimates and Sensitivity Analyses of Ratio of Odds Ratios for Experimental Designs

Comparison	*N*	ROR	95% CI	I^2 [95% CI]	τ^2
HC study compared with summary estimate	25	1.29	[1.01, 1.63]	69 [44, 85]	0.20
Excluding HC study estimates recalculated from primary articles[a]	24	1.30	[1.01, 1.67]	67 [42, 84]	0.22
RRs and HRs substituted for ORs[b]	24	1.34	[1.06, 1.70]	66 [37, 84]	0.19
Excluding coined estimates	17	1.21	[0.97, 1.51]	47 [9, 77]	0.09
HC study is the earliest study	10	1.32	[1.07, 1.62]	0 [0, 61]	0
Meta-analytic comparison mentioned in HC study abstract	21	1.33	[1.02, 1.75]	74 [52, 88]	0.26
HC study compared with largest study	18	2.02	[1.60, 2.57]	17 [0, 61]	0.04
Excluding coined estimates	13	2.18	[1.62, 2.95]	30 [0, 74]	0.09
HC predates largest study	17	2.09	[1.62, 2.68]	18 [0, 62]	0.05
Meta-analytic comparison mentioned in HC abstract	15	2.07	[1.58, 2.72]	32 [0, 73]	0.09

Note: ROR = ratio of odds ratio; CI = confidence interval; HC = highly cited; HR = hazard ratio; RR = risk ratio; OR = odds ratio.
[a]Excluded HC studies: Writing Committee for the ENRICHD Investigators (2003). [b]The comparison excludes cases in which RRs and HRs could not be converted and were considered equivalent to ORs.

the summary ROR estimates became slightly larger when only these topics were considered, a pattern compatible with some influence of "being first." However, the available data are too limited to exclude that this observation may reflect chance.

Finally, our approach of selecting a recent meta-analysis that used emotion-related estimates from the HC studies could have failed to capture the main outcomes of these studies that led to their high citation impact. To account for this possibility, we added a post hoc sensitivity analysis restricted to instances in which the selected meta-analytic comparison included the outcome and exposure/intervention also mentioned in the abstract of the HC study. For most observational (65%) and experimental (85%) HC studies, this was indeed the case, and this sensitivity analysis resulted in very similar results to the main analysis. Of course, the approach cannot fully guarantee we examined the principal finding of the HC study, and it is often impossible to single out only one particular finding from a complex study. However, given that abstracts describe what are considered by the authors to be the most noteworthy results, this approach could represent a useful proxy to identifying the principal findings. We did not assess the quality of the HC studies because this would have posed significant challenges given the diversity of topics, designs, and scientific standards at the time of publication. Although study size is not a surrogate for quality, larger studies are more precise in estimating effects. In general, it was uncommon for HC studies to be also the largest ones.

Some of the HC studies had extremely large effects that also differed tremendously from the respective meta-analyses and largest studies. In the most conspicuous case (RORs of 287 and 297, respectively), Hariri et al. (2002) examined neuroimaging differences in amygdala activation in carriers of the short serotonin-transporter-linked promoter region (5-HTTLPR) allele (one or two copies) compared with those of the long 5-HTTLPR allele. The authors collected 1,769 citations, and the article's standardized effect size in the matching meta-analysis (Munafò et al., 2008) was an incredible SMD of 3.74 (95% CI = [2.51, 4.97]). In contrast, the summary effect in the meta-analysis was considerably smaller (SMD = 0.62, 95% CI = [0.42, 0.82]), similar to the largest study (Hariri et al., 2005; SMD = 0.6, 95% CI = [0.14, 1.06]). The true effect may actually be entirely null. The reason is that this HC study, as well as the other studies in the meta-analysis, depends on a candidate-gene approach, a design that has since been shown as notoriously unreliable (Ioannidis et al., 2011), even more so in neuropsychiatric genetics (Duncan et al., 2019). Moreover, neuroimaging studies are a classic example of a literature replete with small, underpowered studies with high analytical flexibility and often spurious results (Botvinik-Nezer et al., 2020; David et al., 2013; Szucs & Ioannidis, 2020). The proposed association with amygdala activation (Hariri et al., 2002, 2005) would suggest a role of this genetic polymorphism in depression. However, a very large, rigorous genome-wide association study found absolutely no effect for this polymorphism (Border et al., 2019).

RORs were also very large (22 and 29) for a study (Klin et al., 2002) examining differences in visual fixation patterns between autistic males and control subjects

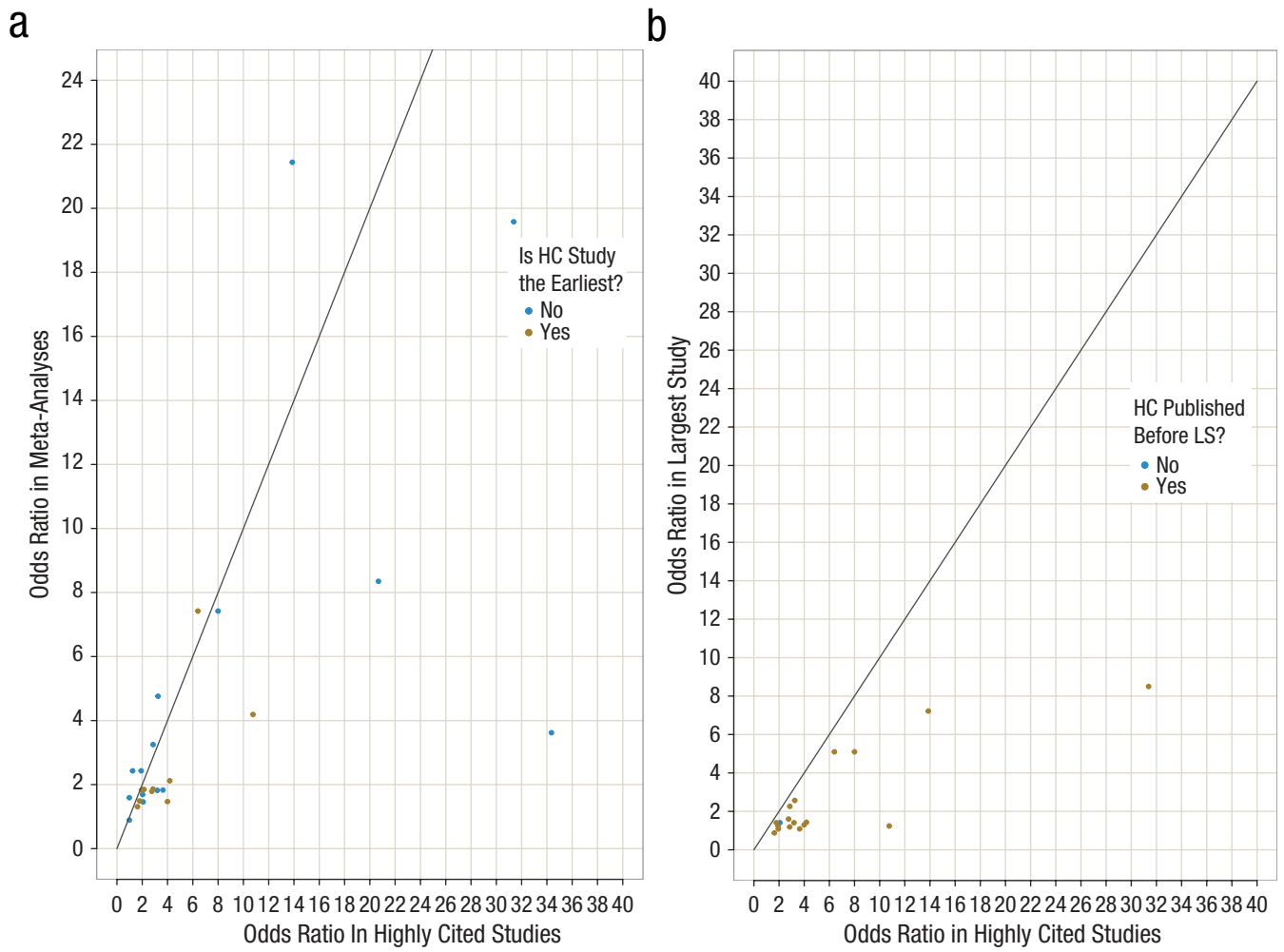

Fig. 3. Scatterplots showing the relation between odds ratios in the highly cited studies and (a) the corresponding summary estimates in the meta-analyses and (b) the corresponding largest-study estimates in experimental designs. The diagonal lines show where the points would fall if the effects were equal. For (b), seven topics in which the highly cited study was the largest study are not shown. LS = largest study.

while viewing social situations (1,150 citations). In this case, the effects between the index study (SMD = −1.47, 95% CI = [−2.27, −0.66]) compared with the corresponding meta-analysis (SMD = 0.24, 95% CI = [0.1, 0.39]) and largest study in it (SMD = 0.39; 95% CI = [0.19, 0.58]) differed not just by magnitude but also by direction. There were no such large outliers in the analyses on experimental studies. Overall, for observational designs, topics in which the original meta-analyses used the SMD as the metric of choice tended to have greater differences in effect size between the HC study and the respective meta-analysis or largest study. Several of the large outliers identified in neuroimaging or genetics belong to this category.

In selecting meta-analyses that included the index study, we focused on the most recent one that contained effect-size data. Around 80% of the identified meta-analyses for observational studies and all but one for experimental studies were published after 2015. The recency of selected meta-analyses makes it more likely that they included a larger number of publications. In addition, the quality of reporting and analysis might also have improved with time (Page et al., 2016; Wen et al., 2008). However, we should caution that the "true" effects for the topic examined are unknown, and effects may genuinely differ across studies on the same topic because of genuine differences rather than bias. Moreover, meta-analyses and even single large studies may also be biased. Random-effects models for obtaining summary results are appropriate in situations in which there is substantial heterogeneity, as is often the case in emotion research, but random-effects estimates are also susceptible to biases such as small-study effects that might underlie publication bias (Sterne et al., 2011). On average, meta-analyses may be more biased than the largest studies. This would be entirely consistent with our observation that HC results seemed to be less inflated when the comparison was made against

the summary effect of a meta-analysis than when it was made against the largest study.

Kvarven et al. (2020) employed a somewhat similar methodological approach to compare results from registered replications with meta-analyses testing the same hypotheses. The starting point were multilaboratory registered replication studies in psychology, for which matching meta-analyses on the same hypothesis, as identified by the study authors, were searched. The authors retrieved meta-analyses with effect-size data for 15 of 62 replication studies selected and used a *Z* test to compare replication effects with summary meta-analysis estimates, either by a random-effects model or with bias adjustment. Results indicated an increase in summary meta-analysis estimates of almost 3-fold compared with replication studies even when using methods to adjust for publication bias, which suggests that better designed studies in which publication bias is avoided (as in the case of preregistered replications) may provide the most accurate effect estimates. If one were to extrapolate from their findings to ours, it is possible that HC studies provide highly inflated results, more inflated than what a comparison against meta-analyses would suggest. Even the comparison against the largest available study may not fully capture the inflation of results because these largest studies that we used were not preregistered. Therefore, they could also suffer from some selective reporting of analyses.

In a study that has direct relevance to the present work, Kvarven et al. (2020) also compared estimates from the original studies—defined as the study that was the object of the replication project—with those in the selected meta-analyses and reported a nonstatistical mean difference of 0.10 for 14 pairs of original studies and meta-analyses. However, the pairs of replication studies and meta-analyses included mostly meta-analyses of small studies, and such meta-analyses may also be unreliable and biased. Likewise, we showed that for meta-analyses with reduced variance, and hence lower uncertainty around the summary effects, differences between estimates from HC studies and summary ones were reduced and no longer significant (RORs close to 1). Conversely, for meta-analyses with higher variance and highly uncertain estimates, differences with HC studies were augmented. In an analysis of 200 meta-analyses published in *Psychological Bulletin*, an eminent journal in psychological science, Stanley et al. (2018) found that only a tiny percentage (< 1%) of experimental studies are adequately powered, compared with about a third of observational studies. Meta-analyses that include only underpowered studies may not be a good "gold standard."

Our findings need to be qualified by important limitations. We were able to identify a matching meta-analysis containing effect-size data for only a third of our sample of target articles. We considered meta-analyses as eligible if they included any emotion-related finding from the target article to avoid ranking findings in the original article in terms of importance. Previous research has dealt with this problem by choosing a finding for which effect-size data are reported in the abstract (Ioannidis & Panagiotou, 2011). However, we were concerned that most of the articles in social and behavioral sciences might simply present findings narratively, with absent or incomplete data, especially in abstracts. Moreover, there is evidence that abstracts are frequently inconsistent with full reports (Li et al., 2017). Nonetheless, ancillary analyses restricted to findings that were mentioned in the abstract supported our main findings. In addition, in the interest of consistency, when a matching meta-analysis included multiple forest plots, we chose the largest one, although it might not have used the most important finding from the HC study. We were able to screen a maximum of 2,000 citations for each target article because of the limitations of exporting data from Scopus. Because we were mostly interested in research on human participants, more general terms such as "fear" or "stress" were not used because they would have rendered the search overtly nonspecific. Finally, we cannot exclude the possibility that in some cases in which effect size was larger in the HC study, the HC study may have been more "correct" than the respective meta-analysis and the largest study on the topic. For instance, the HC study might have had some particularly high-quality features and protection from bias that other studies did not, and biases might have eroded an otherwise genuinely large effect in the other studies. However, this does not seem to be the case in other fields in which HC studies compared with other evidence have been assessed.

Investigations of HC articles in the social and behavioral sciences have been limited and mostly restricted to surveying content and design (Price et al., 2011) or the availability and sharing of the data underlying their findings in HC articles (Hardwicke & Ioannidis, 2018). We add to this metaresearch literature by demonstrating a pervasive systematic citation bias toward exaggerated effects across empirical studies in emotion research.

Transparency

Action Editor: Aidan G. C. Wright
Editor: Jennifer L. Tackett
Author Contributions

I. A. Cristea and J. P. A. Ioannidis were responsible for study concept and design. I. A. Cristea, R. Georgescu, and J. P. A. Ioannidis were responsible for acquisition, analysis, and interpretation of data. I. A. Cristea was responsible for statistical analysis. J. P. A. Ioannidis was responsible for study supervision. I. A. Cristea was responsible for drafting

the manuscript. R. Georgescu and J. P. A. Ioannidis were responsible for critical revision of the manuscript for important intellectual content. All of the authors reviewed and approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

The Meta-Research Innovation Center at Stanford is supported by a grant from the Laura and John Arnold Foundation. The work of J. P. A. Ioannidis is supported by an unrestricted gift from Sue and Bob O'Donnell. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Open Practices

All data have been made publicly available via OSF and can be accessed at https://osf.io/p3x4r. The design and analysis plans for the experiments were preregistered at https://osf.io/wrnq9. This article has received badges for Open Data and Preregistration. More information about the Open Practices badges can be found at https://www.psychologicalscience.org/publications/badges.

ORCID iD

Ioana A. Cristea https://orcid.org/0000-0002-9854-7076

Acknowledgments

We thank Tom E. Hardwicke for contributions to revising the study protocol.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/21677026211049366

References

Altman, D. G., & Bland, J. M. (2003). Interaction revisited: The difference between two estimates. *BMJ*, *326*(7382), Article 219. https://doi.org/10.1136/bmj.326.7382.219

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Border, R., Johnson, E. C., Evans, L. M., Smolen, A., Berley, N., Sullivan, P. F., & Keller, M. C. (2019). No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *American Journal of Psychiatry*, *176*(5), 376–387. https://doi.org/10.1176/appi.ajp.2018.18070881

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., . . . Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite, A., & Poulton, R. (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science*, *301*(5631), 386–389, https://doi.org/10.1126/science.1083968

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*(22), 3127–3131. https://doi.org/10.1002/1097-0258(20001130)19:22<3127::Aid-sim784>3.0.Co;2-m

Cristea, I. A., & Naudet, F. (2018). Defending psychiatry or defending the trivial effects of therapeutic interventions? A citation content analysis of an influential paper. *Epidemiology and Psychiatric Sciences*, *27*(3), 230–239. https://doi.org/10.1017/s2045796017000750

Culverhouse, R. C., Saccone, N. L., Horton, A. C., Ma, Y., Anstey, K. J., Banaschewski, T., Burmeister, M., Cohen-Woods, S., Etain, B., Fisher, H. L., Goldman, N., Guillaume, S., Horwood, J., Juhasz, G., Lester, K. J., Mandelli, L., Middeldorp, C. M., Olié, E., Villafuerte, S., . . . Bierut, L. J. (2018). Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-HTTLPR genotype contributing to the development of depression. *Molecular Psychiatry*, *23*(1), 133–142. https://doi.org/10.1038/mp.2017.44

David, S. P., Ware, J. J., Chu, I. M., Loftus, P. D., Fusar-Poli, P., Radua, J., Munafò, M. R., & Ioannidis, J. P. A. (2013). Potential reporting bias in fMRI studies of the brain. *PLOS ONE*, *8*(7), Article e70104. https://doi.org/10.1371/journal.pone.0070104

Duncan, L. E., Ostacher, M., & Ballon, J. (2019). How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology*, *44*(9), 1518–1523. https://doi.org/10.1038/s41386-019-0389-5

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

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin*, *44*(1), 16–23. https://doi.org/10.1177/0146167217729162

Gerbing, D. (2020). *lessR: Less code, more results* [Computer software] (Version 3.9.8). Comprehensive R Archive Network. https://cran.r-project.org/package=lessR

Gøtzsche, P. C. (1987). Reference bias in reports of drug trials. *British Medical Journal (Clinical Research Edition)*, *295*(6599), 654–656.

Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *BMJ*, *339*, Article 2680. https://doi.org/10.1136/bmj.b2680

Hardwicke, T. E., & Ioannidis, J. P. A. (2018). Populating the data ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLOS ONE*, *13*(8), Article e0201856. https://doi.org/10.1371/journal.pone.0201856

Hariri, A. R., Drabant, E. M., Munoz, K. E., Kolachana, B. S., Mattay, V. S., Egan, M. F., & Weinberger, D. R. (2005). A susceptibility gene for affective disorders and the response of the human amygdala. *Archives of General Psychiatry*, *62*(2), 146–152. https://doi.org/10.1001/archpsyc.62.2.146

Hariri, A. R., Mattay, V. S., Tessitore, A., Kolachana, B., Fera, F., Goldman, D., Egan, M. F., & Weinberger, D. R. (2002). Serotonin transporter genetic variation and the response of the human amygdala. *Science*, *297*(5580), 400–403. https://doi.org/10.1126/science.1071829

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, *294*(2), 218–228. https://doi.org/10.1001/jama.294.2.218

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A., & Panagiotou, O. A. (2011). Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *JAMA*, *305*(21), 2200–2210. https://doi.org/10.1001/jama.2011.713

Ioannidis, J. P. A., Patsopoulos, N. A., & Rothstein, H. R. (2008). Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ*, *336*(7658), 1413–1415. https://doi.org/10.1136/bmj.a117

Ioannidis, J. P. A., Tarone, R., & McLaughlin, J. K. (2011). The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology*, *22*(4), 450–456. https://doi.org/10.1097/EDE.0b013e31821b506e

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowsa, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry*, *59*(9), 809–816. https://doi.org/10.1001/archpsyc.59.9.809

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lai, H. M., Cleary, M., Sitharthan, T., & Hunt, G. E. (2015). Prevalence of comorbid substance use, anxiety and mood disorders in epidemiological surveys, 1990-2014: A systematic review and meta-analysis. *Drug and Alcohol Dependence*, *154*, 1–13. https://doi.org/10.1016/j.drugalcdep.2015.05.031

Li, G., Abbade, L. P. F., Nwosu, I., Jin, Y., Leenus, A., Maaz, M., Wang, M., Bhatt, M., Zielinski, L., Sanger, N., Bantoto, B., Luo, C., Shams, I., Shahid, H., Chang, Y., Sun, G., Mbuagbaw, L., Samaan, Z., Levine, M. A. H., . . . Thabane, L. (2017). A scoping review of comparisons between abstracts and full reports in primary biomedical research. *BMC Medical Research Methodology*, *17*(1), Article 181. https://doi.org/10.1186/s12874-017-0459-5

Martinez-Calderon, J., Flores-Cortes, M., Morales-Asencio, J. M., & Luque-Suarez, A. (2019). Pain-related fear, pain intensity and function in individuals with chronic musculoskeletal pain: A systematic review and meta-analysis. *Journal of Pain*, *20*(12), 1394–1415. https://doi.org/10.1016/j.jpain.2019.04.009

McKinnon, M. C., Yucel, K., Nazarov, A., & MacQueen, G. M. (2009). A meta-analysis examining clinical predictors of hippocampal volume in patients with major depressive disorder. *Journal of Psychiatry & Neuroscience*, *34*(1), 41–54.

Moffitt, T. E., Caspi, A., Harrington, H., & Milne, B. J. (2002). Males on the life-course-persistent and adolescence-limited antisocial pathways: Follow-up at age 26 years. *Development and Psychopathology*, *14*, 179–207. https://doi.org/10.1017/s0954579402001104

Munafò, M. R., Brown, S. M., & Hariri, A. R. (2008). Serotonin transporter (5-HTTLPR) genotype and amygdala activation: A meta-analysis. *Biological Psychiatry*, *63*(9), 852–857. https://doi.org/10.1016/j.biopsych.2007.08.016

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Odgers, C. L., Moffitt, T. E., Broadbent, J. M., Dickson, N., Hancox, R. J., Harrington, H., Poulton, R., Sears, M. R., Thomson, W. M., & Caspi, A. (2008). Female and male antisocial trajectories: From childhood origins to adult outcomes. *Development and Psychopathology*, *20*(2), 673–716. https://doi.org/10.1017/s0954579408000333

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Page, M. J., Shamseer, L., Altman, D. G., Tetzlaff, J., Sampson, M., Tricco, A. C., Catalá-López, F., Li, L., Reid, E. K., Sarkis-Onofre, R., & Moher, D. (2016). Epidemiology and reporting characteristics of systematic reviews of biomedical research: A cross-sectional study. *PLOS Medicine*, *13*(5), Article e1002028. https://doi.org/10.1371/journal.pmed.1002028

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*(9), 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Paule, R. C., & Mandel, J. (1989). Consensus values, regressions, and weighting factors. *Journal of Research of the National Institute of Standards and Technology*, *94*(3), 197–203. https://doi.org/10.6028/jres.094.020

Polanin, J. R., & Snilstveit, B. (2016). Converting between effect sizes. *Campbell Systematic Reviews*, *12*(1), 1–13. https://doi.org/10.4073/cmpn.2016.3

Price, K. W., Floyd, R. G., Fagan, T. K., & Smithson, K. (2011). Journal article citation classics in school psychology: Analysis of the most cited articles in five school psychology journals. *Journal of School Psychology*, *49*(6), 649–667. https://doi.org/10.1016/j.jsp.2011.10.001

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.0) [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org

Regier, D. A., Farmer, M. E., Rae, D. S., Locke, B. Z., Keith, S. J., Judd, L. L., & Goodwin, F. K. (1990). Comorbidity of mental disorders with alcohol and other drug abuse. Results from the Epidemiologic Catchment Area (ECA) Study. *JAMA*, *264*(19), 2511–2518.

Reising, K., Ttofi, M. M., Farrington, D. P., & Piquero, A. R. (2019). Depression and anxiety outcomes of offending trajectories: A systematic review of prospective longitudinal studies. *Journal of Criminal Justice*, *62*, 3–15. https://doi.org/10.1016/j.jcrimjus.2018.05.002

Richards, S. H., Anderson, L., Jenkinson, C. E., Whalley, B., Rees, K., Davies, P., Bennett, P., Liu, Z., West, R., Thompson, D. R., & Taylor, R. S. (2017). Psychological interventions for coronary heart disease. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD002902.pub4

RStudio Team. (2019). *RStudio: Integrated development for R* [Computer software] (Version 1.2.5033). http://www.rstudio.com/

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

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

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*, Article d4002. https://doi.org/10.1136/bmj.d4002

Swendsen, J. D., Merikangas, K. R., Canino, G. J., Kessler, R. C., Rubio-Stipec, M., & Angst, J. (1998). The comorbidity of alcoholism with anxiety and depressive disorders in four geographic communities. *Comprehensive Psychiatry*, *39*(4), 176–184. https://doi.org/10.1016/S0010-440X(98)90058-X

Szucs, D., & Ioannidis, J. P. A. (2020). Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990-2012) and of latest practices (2017-2018) in high-impact journals. *NeuroImage*, *221*, Article 117164. https://doi.org/10.1016/j.neuroimage.2020.117164

Tajika, A., Ogawa, Y., Takeshima, N., Hayasaka, Y., & Furukawa, T. A. (2015). Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up. *British Journal of Psychiatry*, *207*(4), 357–362. https://doi.org/10.1192/bjp.bp.113.143701

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*(1), 55–79. https://doi.org/10.1002/jrsm.1164

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, *26*(1), 37–52. https://doi.org/10.1002/sim.2514

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Blouin-Hudon, E. M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., Dijkhoff, L., Dijkstra, K., Fischer, A. H., . . . Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. https://doi.org/10.1177/1745691616674458

Wen, J., Ren, Y., Wang, L., Li, Y., Liu, Y., Zhou, M., Liu, P., Ye, L., Li, Y., & Tian, W. (2008). The reporting quality of meta-analyses improves: A random sampling study. *Journal of Clinical Epidemiology*, *61*(8), 770–775. https://doi.org/10.1016/j.jclinepi.2007.10.008

Writing Committee for the ENRICHD Investigators. (2003). Effects of treating depression and low perceived social support on clinical events after myocardial infarction: The Enhancing Recovery in Coronary Heart Disease Patients (ENRICHD) randomized trial. *JAMA*, *289*(23), 3106–3116. https://doi.org/10.1001/jama.289.23.3106