nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-019-0787-z

Comparing meta-analyses and preregistered multiple-laboratory replication projects

Amanda Kvarven[1,3], Eirik Strømland[1,3] and Magnus Johannesson[2*]

Many researchers rely on meta-analysis to summarize research evidence. However, there is a concern that publication bias and selective reporting may lead to biased meta-analytic effect sizes. We compare the results of meta-analyses to large-scale preregistered replications in psychology carried out at multiple laboratories. The multiple-laboratory replications provide precisely estimated effect sizes that do not suffer from publication bias or selective reporting. We searched the literature and identified 15 meta-analyses on the same topics as multiple-laboratory replications. We find that meta-analytic effect sizes are significantly different from replication effect sizes for 12 out of the 15 meta-replication pairs. These differences are systematic and, on average, meta-analytic effect sizes are almost three times as large as replication effect sizes. We also implement three methods of correcting meta-analysis for bias, but these methods do not substantively improve the meta-analytic results.

Meta-analyses are often placed at the top of the hierarchy of scientific evidence[1]. The quantitative summary provided by a meta-analysis makes it easier to navigate through large scientific literatures, and meta-analyses have far greater statistical power than individual studies. For these reasons, meta-analysis is viewed as an attractive tool for summarizing scientific research[1–3]. In the past 30 years, the number of meta-analyses published across scientific fields has been growing exponentially[4] and some scholars have called for greater reliance on 'meta-analytic thinking' in the behavioural sciences[2].

However, the properties of a meta-analysis depend on the primary studies that it includes; if primary studies overestimate effect sizes in the same direction, so too will the meta-analysis. The recent surge of replication studies in the behavioural sciences suggests that original studies produce larger effect sizes than replication studies[5–10]. The pioneering Reproducibility Project: Psychology (RP:P) replicated 100 studies in psychology and found that effect sizes standardized to the correlation coefficients (r) were on average about twice as large in the original studies as in the replication studies[10]. Similar relative effect sizes were reported by the Experimental Economics Replication Project and the Social Sciences Replication Project[5,6].

The substantially higher effect sizes in the original studies compared to the replications are likely, at least partially, to be caused by publication bias and selective reporting of statistically significant results[11–22]. We use the term 'publication bias' to refer to bias due to the behaviour of journals, and the term 'selective reporting' to refer to bias due to the behaviour of researchers such as selective outcome reporting, *P*-hacking, significance chasing, garden-of-forking paths, misreporting of results and researcher degrees of freedom. For examples of direct and indirect evidence on publication bias and selective reporting see, for instance, Simmons, Nelson and Simonsohn[14], Brodeur[17], John, Loewensten and Prelec[20] and Franco, Malhotra and Simonovits[21,22]. In line with these biases, systematic reviews of meta-analyses, sometimes referred to as 'meta-meta-analysis', suggest that smaller studies are associated with larger effect sizes than larger studies (the small study effect) and that unpublished studies are associated with smaller effect sizes than published studies[23–26].

Because problems with selective reporting translate into biased meta-analytic effect sizes, some have argued that reliance on meta-analysis will exacerbate the problems of publication bias and selective reporting, and that greater reliance on meta-analytic thinking in the behavioural sciences will increase the rate of false positives[27,28]. Others argue that meta-analyses could reduce the influence of publication bias and help improve reproducibility in the behavioural sciences[2]. Several meta-analytic methods have been developed that aim to adjust effect sizes for the influence of publication bias[11,29–32], but simulation studies typically find that different selection methods can perform either very well or poorly depending on the particular setting[33–36]. In this study, we provide empirical evidence on the ability of both standard meta-analysis and adjustment methods to produce unbiased effect sizes and accurate summary conclusions based on evidence from primary studies.

Our approach is to use large-scale registered replication studies in psychology carried out at multiple laboratories—where publication bias and selective reporting of results are eliminated by construction—as a baseline to which the results of meta-analyses on the same topics will be compared. We refer to these studies as replication studies and replication effect sizes below, and contrast these with meta-analysis studies and meta-analytic effect sizes. We refer to the study being replicated as the original study. We focus on multiple-laboratory replications, as these involve large sample sizes leading to relatively precisely estimated effect sizes. This increases the statistical power of finding a significant difference between replication and meta-analytic effect sizes. Multiple-laboratory replication studies are in themselves also meta-analyses, as they use meta-analysis to pool effect sizes across laboratories. Using our methodology, we can both estimate the fraction of studies for which replication and meta-analytic effect sizes differ significantly, and estimate to what extent these differences are systematic. Moreover, our methodology can be applied to test the performance of different methods for bias adjustment proposed in the literature.

[1]Department of Economics, University of Bergen, Bergen, Norway. [2]Department of Economics, Stockholm School of Economics, Stockholm, Sweden. [3]These authors contributed equally: Amanda Kvarven, Eirik Strømland. *e-mail: magnus.johannesson@hhs.se

Results

Meta-replication pairs found in the literature search. We started by collecting data on studies in psychology where many different laboratories joined forces to replicate a well-known effect according to a pre-analysis plan. We identified two data sources in line with this criterion: (1) replications published according to the 'registered replication report' format in the journals 'Perspectives on Psychological Science' and 'Advances in Methods and Practices in Psychological Science'[37]; and (2) The 'Many Labs' projects in psychology[7–9]. Both data sources feature a set of independent laboratories replicating some original study according to a pre-defined analysis plan. We identified 62 such replication effects (see Methods for details).

After identification of relevant replication experiments, we searched for meta-analyses on the same research question. An initial search identified 39 studies of potential interest that were further assessed for eligibility (see Supplementary Table 1). Of these 39 meta-analyses, 21 were excluded due to a lack of correspondence in the effects estimated in the meta-analyses and replication studies, and two were excluded due to lack of data. Of the remaining 16 meta-analyses, two studied the same effect. To ensure that our observations were statistically independent, we chose to include the largest meta-analysis in the main analysis and include the other in a robustness test. One of remaining 15 meta-analyses had three separate estimates that could be matched to the same replication estimate, and we therefore selected the most precise of these for inclusion in the main analysis and used the other two in robustness checks (see Methods for details on the inclusion of studies and Fig. 1 for a flow diagram; PRISMA is an evidence-based minimum set of items used for reporting in systematic reviews and meta-analyses). In total we thus included 15 replication meta-pairs in our baseline analysis, and our final dataset spans 15 preregistered replication studies using a multiple-laboratory format ($n = 53{,}365$ subjects in total) and 15 corresponding meta-analyses on the same research question ($n = 336{,}027$ subjects in total; see Supplementary Tables 2–4 for details of the 15 original studies[38–52], meta-analyses and replications).

Of these 15 studies, 11 meta-analyses included the original study replicated in the replication studies and one meta-analysis included the replication study. For the 11 meta-analyses that included the original study, our judgement and those of the original meta-analysts coincided by definition and we provide a robustness test using this sub-sample below (see Methods for a discussion on the four meta-analyses that did not include the original study).

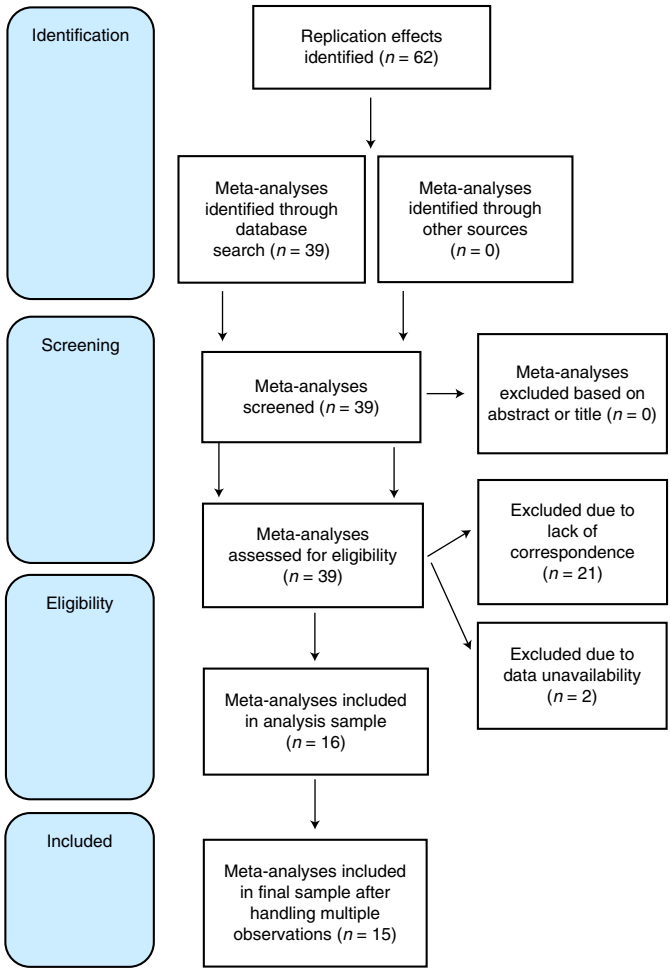

Fig. 1 | PRISMA flow diagram showing the number of meta-analyses considered for inclusion. The diagram illustrates our process of data collection, leading up to our analysis sample of 15 meta-replication pairs. *n*, number of observations. Note that the flow diagram indexes *n* as the number of multiple-laboratory replication studies in the first box and as the number of meta-analyses in subsequent boxes, not the number of studies included in the meta-analysis, which is the conventional use of this flow diagram in meta-analysis studies.

Comparison of meta-analytic and replication effect sizes. We converted all effect sizes to Cohen's *d*, with the exception of one meta-replication pair[41]. One of the replication studies measured effect sizes in Cohen's *q* units, and there is no established way of converting effect sizes from Cohen's *q* to *d*. We therefore measured effect sizes in Cohen's *q* for this meta-replication pair and treated it as being equivalent to Cohen's *d* in the analysis, but also performed a robustness test without this meta-replication pair.

We compared the meta-analytic to the replication effect size for each effect using a *z*-test (see Methods for details). We used a *z*-test to determine whether the mean effect size differed between the replication study and the meta-analysis, and this test was based on the reported mean effect and standard error from both the replication study and the meta-analysis. We wanted to use these estimates as reported in the replication and meta-analysis papers, to compare meta-analysis as practised in multiple-laboratory replication studies. All our statistical tests are two-tailed and follow the recent recommendation to refer to tests with $P < 0.005$ as statistically significant and $P = 0.05$ as suggestive evidence against the null[53]. In Supplementary Tables 5 and 6 we show with what effect size difference we have 80% power to detect at the 5% (suggestive evidence) and the 0.5% (statistically significant evidence) levels, for each of the 15 meta-replication pairs and for the pooled overall difference for the 15 studies.

In Fig. 2 we show the 95 and 99.5% confidence intervals (CIs) of the meta-analytic and replication effect size for each study pair (Fig. 2), and we show the CIs of the difference in effect size for each study pair (Fig. 2) (see also Supplementary Table 5 for detailed results). The direction of effect size is based on the direction of the effect reported in the original study that was replicated (a positive effect implies an effect in the same direction as the original study, and a negative effect implies an effect in the opposite direction).

As seen in Fig. 2a, the meta-analysis and replication studies reach the same conclusion about the direction of the effect using the 0.005 statistical significance criterion for seven (47%) study pairs; in all seven cases both the meta-analysis and replication studies find a significant effect in the same direction as the original study. For seven (47%) study pairs, the meta-analysis finds a significant effect in the original direction whereas the replication cannot reject the null hypothesis and, in the remaining study pair, the meta-analysis cannot reject the null hypothesis whereas the replication study finds

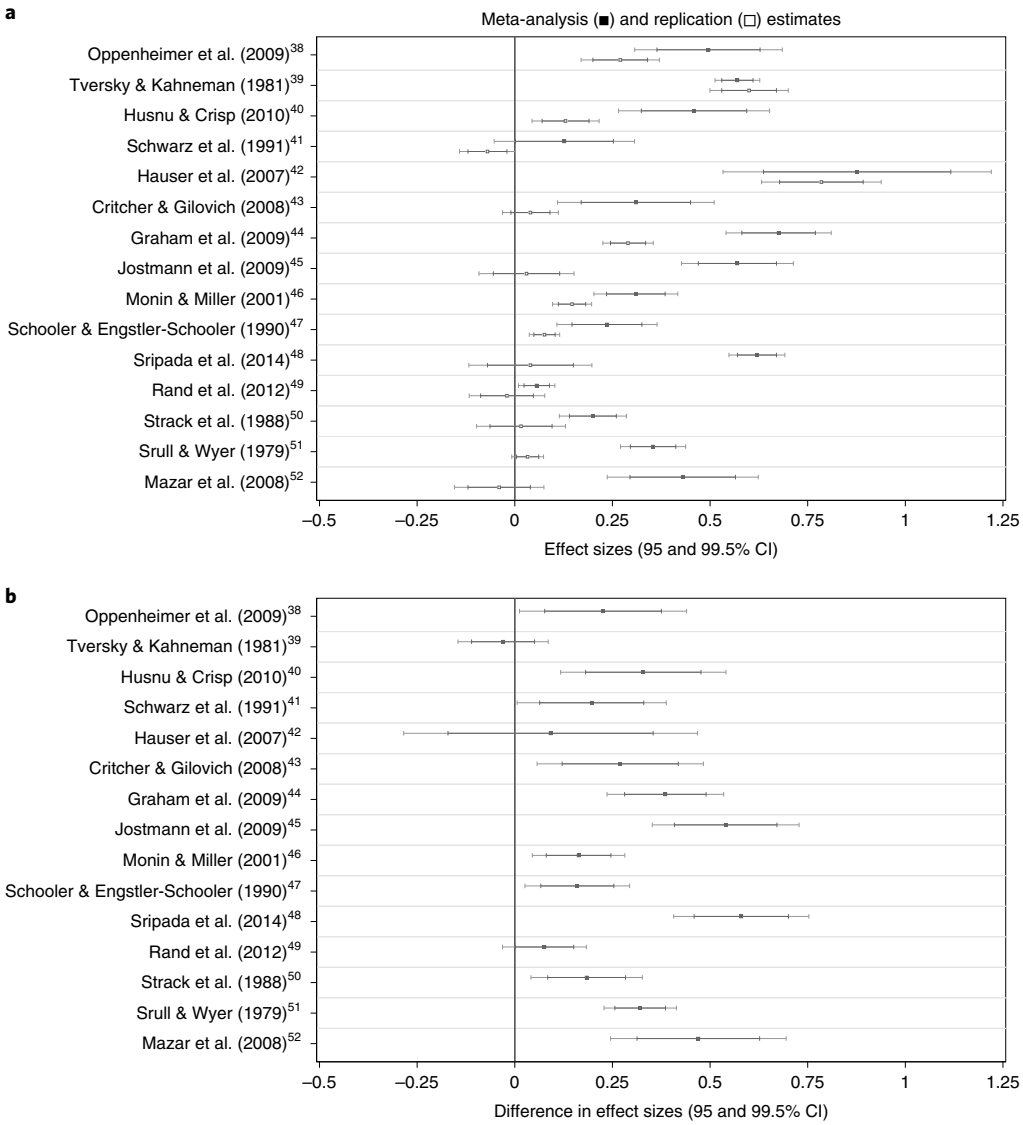

Fig. 2 | Results of meta-analyses and replication studies. a, Plotted are 95 and 99.5% CIs of meta-analytic and replication studies effect sizes for each study pair estimating the same effects (effect sizes are measured in Cohen's *d*). The references listed for the 15 studied effects are the 15 original studies replicated in the replication studies. **b**, Plotted are 95 and 99.5% CIs of the difference in meta-analytic and replication studies effect size for each study pair estimating the same effects.

a significant effect in the opposite direction to the original study. Note that the replication power is high for the Many Labs replication projects (see Supplementary Table 4), and failing to reject the null hypothesis for these replication studies is unlikely to be due to insufficient power. Power is also generally high for the meta-analyses because most have sufficient power to detect a small effect, although three meta-analyses have 80% power to detect a medium effect at the 0.5% level (see Supplementary Table 3). In Fig. 2b we can see that the difference in estimated effect size is significant for 12 (80%) of the studies, and there is suggestive evidence of a difference for one additional study. For all 12 studies, the effect size is higher in the meta-analysis. For some of the meta-replication pairs the power to find a significant difference is limited, as indicated by the wide variation in CIs for effect size difference (see also Supplementary Table 5).

The observed rate of significant differences in effects sizes between meta-analyses and replication studies is high, and this pattern is reinforced by comparing average effect sizes between the two studies. The average unweighted effect size is 0.155 for the 15 replication studies and 0.419 for the 15 meta-analysis studies, implying that the mean meta-analytic effect size is almost three times as large as the mean replication effect size. To further estimate to what extent there are systematic differences in average effect size between studies, we used random effects meta-analysis to estimate the mean effect size difference across the 15 study pairs in our sample (see Methods for details). This analysis approach could be thought of as a meta-meta-analysis. In Fig. 3 we show the CIs for the mean difference in replication and meta-analytic effect size. The estimated mean difference is 0.263 and is highly significant ($n = 15$, $z = 5.810$, $P < 0.001$, 95% CI = 0.175–0.352, 99.5% CI = 0.136–0.391). A non-parametric Wilcoxon test on the 15 paired meta-replication differences produced the same outcome ($n = 15$, $W = 1$, $P < 0.001$).

Robustness tests and sub-group analyses. We also conducted several robustness tests and sub-group analyses of differences in mean meta-analytic and mean replication effect size in the meta-meta-analysis (Fig. 3; see also Supplementary Table 6 for detailed results). We performed three robustness tests on the inclusion criteria:

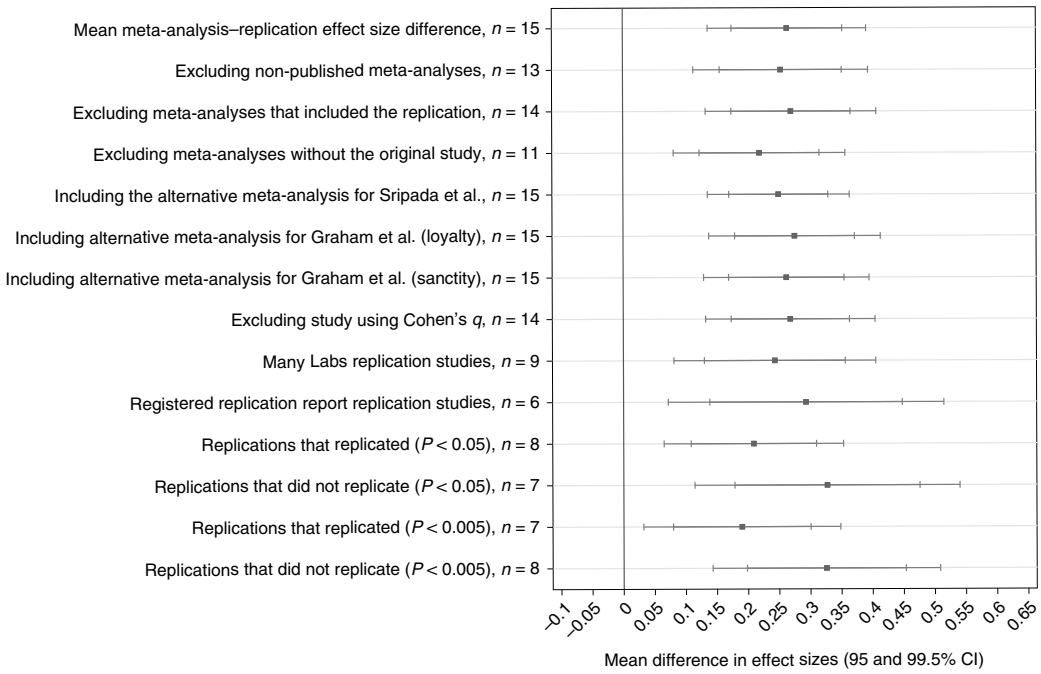

Fig. 3 | Mean effect size difference across the 15 meta-replication pairs in our sample, and robustness test and sub-group analyses of this difference. Random effects meta-analysis was used to estimate the mean effect size difference. For each analysis we plot the 95 and 99.5% CIs of mean difference in the meta-analytic and replication effect sizes for all effects included in that analysis (effect sizes measured in Cohen's *d*). The top row is our main random effects estimate using the entire analysis sample of 15 meta-replication pairs.

excluding unpublished meta-analyses ($n=2$), so our analysis sample is $n=13$; excluding meta-analyses that included the replication study ($n=1$), so our analysis sample is $n=14$; and excluding meta-analyses ($n=4$) that did not include the original study that was replicated, so our analysis sample is $n=11$. We also carried out a robustness test excluding the meta-replication pair with effect sizes measured in Cohen's *q*, so our sample is $n=14$. In an additional robustness test, we also used the alternative meta-analysis for the replication study where we identified two meta-analyses. In two final robustness checks we used the two alternative meta-analysis estimates reported in the same meta-analysis study corresponding to the original study[44]. The results derived are similar to the main results in these robustness tests. This is perhaps not surprising, as there is very large overlap between the data in each of these robustness tests and those in the main analysis based on the 15 meta-replication pairs.

We also carried out some sub-group analyses. We report results separately for replications from the Many Labs projects ($n=9$) and replications from registered replication report studies ($n=6$), because the selection of studies for replication can differ between these. The results are similar in these two sub-groups as well, although with a slightly higher point estimate of the difference for the registered replication report studies.

Finally, we separated results into sub-groups depending on whether the effect was significant in the replication study; we defined these groups based on a significance threshold of both $P<0.005$ and $P<0.05$. This was done to test whether the difference between the meta-analytic and replication effect size is driven by studies where the null hypothesis cannot be rejected in the replication. If the null hypothesis is true, selective publication may not necessarily result in biased meta-analytic effect sizes if 'significant' results with positive and negative signs cancel out. However, Nelson, Simmons and Simonsohn[27] and Vosgerau et al.[28] suggest that meta-analyses are prone to producing false positives, and that aggregation in meta-analyses in general does not lead to cancelling-out of errors. Supporting this mechanism, we find a significant and large difference measure between meta-analytic and replication effect size for replication studies that cannot reject the null hypothesis. Although the point estimates are smaller, the difference measure is significant also for replication studies rejecting the null hypotheses, implying that effect size inflation of both true hypotheses and false positives may have contributed to our results. We furthermore tested whether the difference measure is significantly smaller for replication studies rejecting the null hypothesis than for those failing to reject it, but found no evidence for this (see Supplementary Table 7).

To summarize our findings, we find that there is a significant difference between meta-analytic and replication effect size for 12 of the 15 studies (80%), and suggestive evidence for a difference in one additional study. These differences are systematic—the meta-analytic effect size is larger than the replication effect for all these studies—and on average for all 15 studies the estimated effect sizes are almost threefold higher in the meta-analyses. However, this point estimate of the degree of overestimation should be interpreted cautiously because the size of overestimation varies considerably across studies—from no overestimation to an overestimation of >0.5 Cohen's *d*. Interestingly, the relative difference in estimated effect size is of at least the same magnitude as that observed between replications and original studies in the RP:P and other similar systematic replication projects[5,6,10]. Publication bias and selective reporting in original studies have been suggested as possible reasons for the low reproducibility in RP:P and other replication projects, and our results imply that these biases are not eliminated by the use of meta-analysis. This is not surprising if it is the case that the same publication biasing factors are at work in non-original studies of a phenomenon as in the original studies, and meta-analyses are unsuccessful at including unpublished studies. However, in general the direction of the biases in meta-analyses depends on a host of often unknown factors. In many cases the meta-analysis might focus

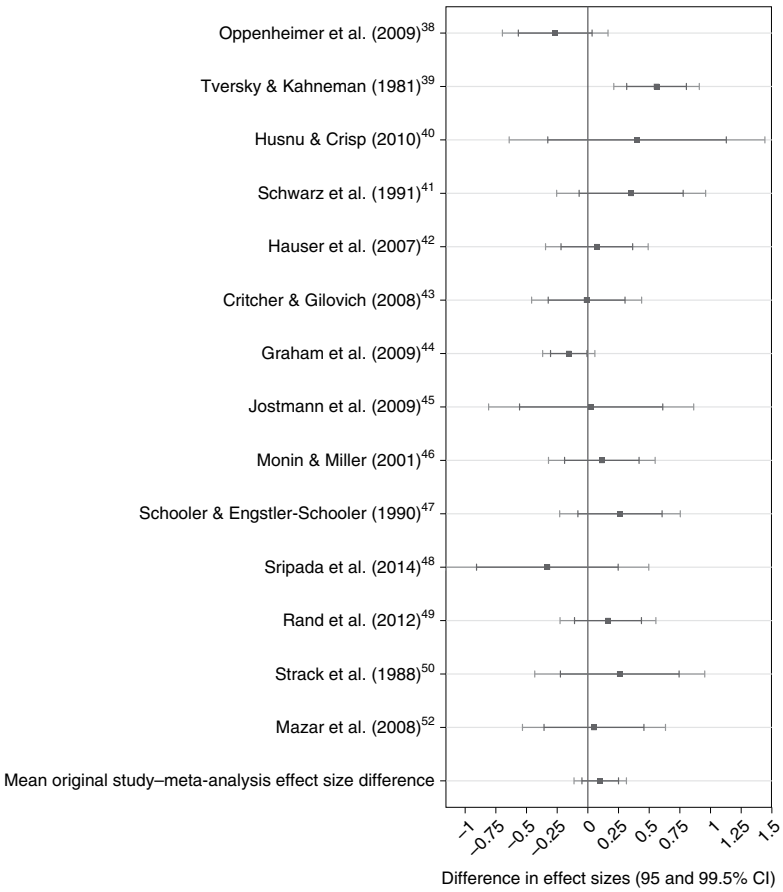

Fig. 4 | Comparison of effect sizes of the original studies replicated in the replication studies to meta-analytic effect sizes. Plotted are 95 and 99.5% CIs of the differences in original study and meta-analytic effect sizes for each original study–meta-analysis pair estimating the same effect (effect sizes are measured in Cohen's *d*). We lack information about the effect size in Cohen's *d* for the original study of Srull and Wyer[51], and therefore the difference in the original study and meta-analytic effect size is shown for only 14 effects. The 95 and 99.5% CIs of the mean effect size difference across the 14 original study and meta-analysis pairs in our sample are plotted, estimated using a random effects model. The references listed for the 14 included effects are the 14 original studies replicated in the replication studies.

on an effect where the authors have no vested interests, thereby lowering potential biases. In other cases, the meta-analysis may focus on an effect where the authors have a strong vested interest and are hence potentially subject to biases. Previous meta-meta-analyses have also investigated several meta-scientific biases, including the decline effect and the early-extreme effect[54].

Heterogeneity in meta-analyses. In a meta-analysis there can be heterogeneity in the observed effects due to variations in the true effect size among different populations (sample heterogeneity) and different designs used to test the hypothesis (design heterogeneity). By 'design heterogeneity' we mean any difference in how a hypothesis is tested between studies apart from a difference in samples. In the multiple-laboratory replication studies included in our study, the design is held constant across laboratories whereas the samples vary. In the meta-analyses both sample and design heterogeneity can lead to variation in effect size, which suggests that there will be greater heterogeneity in effect size for the meta-analyses than for replication studies. However, a recent study suggests that publication and related biases can have complex effects on study heterogeneity[55].

Higher heterogeneity per se in the meta-analyses cannot explain our findings, but higher heterogeneity increases the potential for replication studies to select samples or designs associated with systematically lower true effect sizes. We refer to this potential mechanism as 'replicator selection'.

For sample heterogeneity to explain our results, the replications need to have been conducted in samples with, on average, lower true effect sizes than in those included in the meta-analyses. This explanation seems implausible in our setting, where the replication studies consist of multiple-laboratory studies in different samples that are pooled. The Many Labs replication studies also suggest that sample heterogeneity is not sufficiently large to potentially explain our findings[7–9]. These studies often report no significant between-study heterogeneity[7–9] and, in the recent Many Labs 2 study, the reported standard deviation in the true effect size across labs (Tau) was 0 for 19 of the 28 studies while the average was 0.04 (ref. [8]).

For design heterogeneity to explain our results, replication studies must select experimental designs producing lower true effect sizes than the average design used to test the same hypotheses in the meta-analyses. The design heterogeneity would have to be substantial and the replicator selection of weak designs strong for this mechanism to explain our findings. Replicator selection implies a positive correlation between heterogeneity in the meta-analysis and the observed difference in meta-analytic and replication effect sizes, because larger heterogeneity increases the scope for replicator selection. To test this, we computed the standard deviation in true effect sizes across studies (Tau) for the meta-analyses in our sample

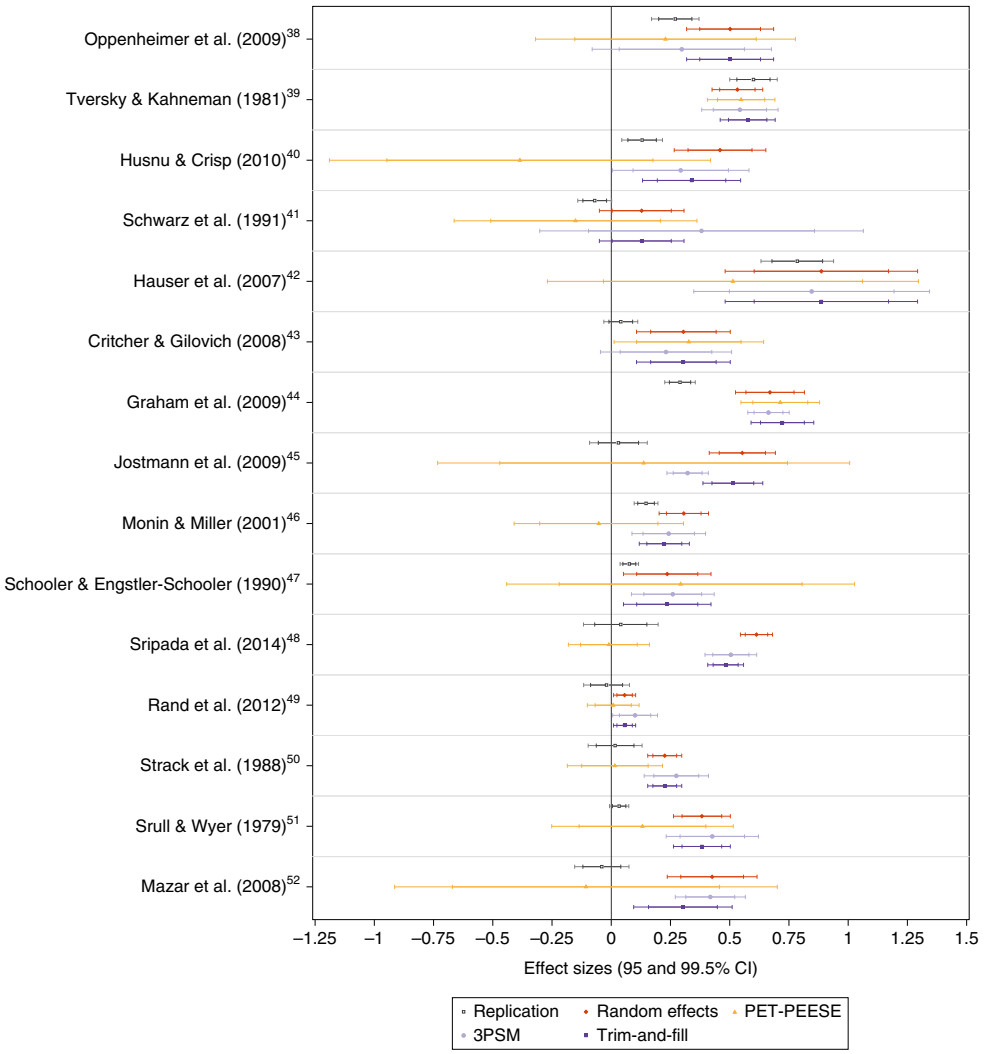

Fig. 5 | Results for methods of correcting meta-analyses for bias. Plotted are 95 and 99.5% CIs of replication and meta-analytic effect sizes estimated with the random effects model and the three bias-adjustment methods (trim-and-fill, PET-PEESE and 3PSM) for each study pair estimating the same effects (effect sizes are measured in Cohen's *d*). The random effects model is used as a benchmark for intercomparison of the three bias-adjustment methods. The references listed for the 15 studied effects are the 15 original studies replicated in the replication studies.

(see Supplementary Table 3), and estimated the correlation between Tau and the difference in meta-analytic and replication effect sizes. Mean Tau was 0.305, ranging from 0 to 0.735, the Spearman correlation was −0.300 ($n=15$, $P=0.277$) and the Pearson correlation −0.406 ($n=15$, $P=0.133$). These correlations show no indication of replicator selection in our data, and the point estimates are in the opposite direction to that predicted by the replicator selection mechanism. This finding contrasts with a recent paper that attributes reproducibility failures in psychology to heterogeneity in the underlying effect sizes[56].

To further test for replicator selection, we compared the original to the meta-analytic effect sizes. If the original study and the other studies in the meta-analysis are equally affected by selective reporting and publication bias, replicator selection implies a lower effect size in the original study than in the others. We were able to obtain effect sizes and standard errors of the original studies converted to Cohen's *d* for all original studies except one, where the standard deviation was unavailable (see Supplementary Table 2)[51]. In Fig. 4 we show the confidence intervals of the difference in the original and the meta-analytic effect size for these 14 observations (based on *z*-test as above). The replicator selection hypothesis implies that the effect size should be lower in the original studies, but this is the case only in four observations and the difference is small and non-significant in these observations (for one of these there is suggestive evidence for a difference). For the other ten observations the difference trends in the opposite direction, with higher effect sizes in the original studies, but only one of those differences is significant (see Supplementary Table 8 for details). The average unweighted effect size of these 14 original studies is 0.531 compared to an average unweighted effect size of 0.424 of the same 14 studies in the meta-analyses. The mean random effects difference in original and meta-analytic effect sizes is 0.09, but this difference is not statistically significant (see Supplementary Table 8 for details). We therefore do not find support for the replicator selection hypothesis in this analysis either. If anything, the comparison suggests that we may have underestimated the difference in meta-analytic and replication effect size. However, the comparison in original and meta-analytic effect size should be interpreted very cautiously as the assumption of a constant bias due to selective reporting and publication bias is a strong one. One could, alternatively, interpret the comparison of original and meta-analytic effect size as a test of whether meta-analyses reduce the influence of publication bias or selective reporting

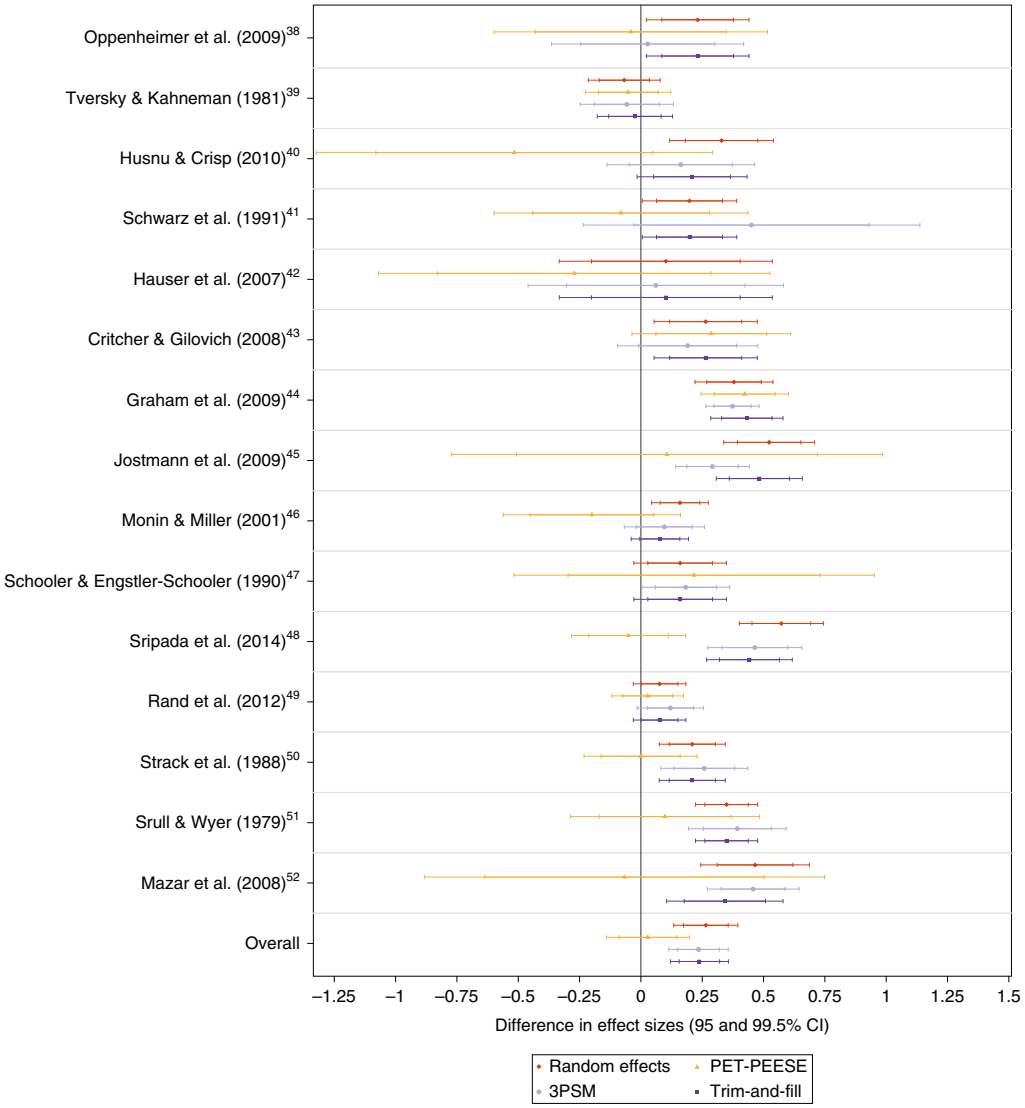

Fig. 6 | Estimated differences in meta-analytic and replication studies effect sizes for methods of correcting meta-analyses for bias. Plotted are 95 and 99.5% CIs of the difference in meta-analytic and replication studies effect sizes for each study pair estimating the same effects (effect sizes are measured in Cohen's *d*). Differences are shown for four meta-analytic models—the benchmark random effects model and the three bias-adjustment methods (trim-and-fill, PET-PEESE and 3PSM). The final row plots the 95 and 99.5% CIs of the mean effect size difference across the 15 meta-replication pairs in our sample estimated with a random effects model. The references listed for the 15 studied effects are the 15 original studies replicated in the replication studies.

compared to the original studies, and with that interpretation the point estimate of the random effect difference is consistent with meta-analyses reducing the inflated effect sizes in original studies somewhat (although the effect is not statistically significant).

Evaluating meta-analysis methods for bias adjustment. Our results suggest that the effect sizes reported in the 15 meta-analyses substantially overestimate the true effect sizes; given this fact, it is interesting to ask whether it is possible to adjust meta-analytic effect sizes for overestimation. Because publication bias is a well-known threat to the validity of meta-analysis, there exist various tests of publication bias and several different bias-adjustment methods and estimators for correction of publication bias[11,29,31,34,35,57]. Simulation studies suggest that these bias-adjustment methods often fail to adjust for publication bias, or even sometimes lead to underestimation of effect sizes[34,58]. To assess the performance of bias-adjustment methods in our data, we implemented three bias-adjustment methods and compared the results produced by these methods to conventional random effects meta-analysis. We use this form of meta-analysis as a benchmark model for uncorrected meta-analysis, because the bias-adjustment methods are based on the random effects model. The results for random effects meta-analyses differ slightly from those reported in Fig. 1 for some of the meta-analyses, as not all of the original meta-analyses used the random effects model to pool the results of the individual studies included in the meta-analysis. The average effect size (standard error) reported in the 15 original meta-analyses is 0.419 (0.051), compared to 0.419 (0.055) if random effects meta-analysis is used in all 15 meta-analyses.

We restrict our focus to three bias-adjustment methods: (1) the trim-and-fill method developed by Duval and Tweedie[11], as this is the method most commonly used to adjust for publication bias[33]; (2) the PET-PEESE estimator proposed by Stanley and Doucouliagos[31]; and (3) the three-parameter selection model (3PSM) proposed by Hedges and Vevea[29]. 3PSM has been recommended as the minimal model to be considered in applied work in a recent review paper

Table 1 | Results for different indicators used to assess performance of the three meta-analysis bias-adjustment methods

Method	False-positive rate, 0.5% level (5% level)	False-negative rate, 0.5% level (5% level)	Mean meta-replication difference, random effects (*z*-statistic; *P* value)	Mean meta-replication difference, random effects 99.5% CI (95% CI)	Mean meta-replication difference, unweighted	Overestimation factor	Root mean squared error	Mean MDE 0.5% level (5% level)
Random effects	100% (100%)	0% (0%)	0.265 (5.69; <0.001)	0.13, 0.40 (0.17, 0.36)	0.26	2.7	0.31	0.200 (0.16)
PET-PEESE	14.2% (16.6%)	71.4% (75%)	0.0285 (0.47; 0.636)	−0.14, 0.20 (−0.09, 0.15)	−0.01	0.95	0.22	0.60 (0.46)
3PSM	85.7% (100%)	14.3% (0%)	0.235 (5.44; <0.001)	0.11, 0.36 (0.15, 0.32)	0.23	2.49	0.28	0.30 (0.23)
Trim-and-fill	100% (100%)	0% (0%)	0.24 (5.66; <0.001)	0.12, 0.36 (0.16, 0.32)	0.24	2.53	0.28	0.20 (0.16)

In estimating the indicators, the meta-analytic results for the bias-adjustment methods were compared to results for the replication studies (we also included results for the random effects model as a benchmark of the value of the indicators for uncorrected meta-analysis). Results for several indicators are shown for both the 0.5 and 5% significance levels (see Methods for definition of indicators).

by McShane, Böckenholt and Hansen[35], and has been shown to perform relatively well under a large set of conditions in simulation studies[34,35] (see Methods for details on these methods and how we implemented them). Note that valid interpretation of our results for the bias-adjustment methods relies on the assumption that the meta-analyses and replications are comparable, in the sense that they belong to the same distribution of true effects.

Figures 5 and 6 show that the trim-and-fill and 3PSM models generally produce results similar to random effects meta-analysis; the degree of overestimation of effect size is almost identical for these methods, and the pooled effect size difference is statistically significant for all these meta-analytic models (see also Supplementary Tables 9–16 for detailed results). PET-PEESE reduces mean difference substantially, but it also reduces statistical power considerably as seen by the wide CIs. As a consequence, PET-PEESE fails to reject the null more often than the other adjustment methods—the estimated false-negative rate for PET-PEESE is 71.4% for significance level $P=0.005$ and 75% for $P=0.05$, whereas the three other methods do not fail to reject the null when the replication rejects it; the estimated false-negative rate is always 0% for random effects and trim-and-fill, while for 3PSM it is 0 and 14.3% for the $P=0.005$ and $P=0.05$ levels, respectively. Although PET-PEESE results in less overestimation of effect size on average, it does not substantially reduce the average prediction error (root mean squared error) compared to the other bias-adjustment methods (see Table 1 for a detailed comparison of different indicators used for assessment of the performance of the bias-correction methods).

Discussion

A central caveat in interpreting our findings is the potential impact of heterogeneity in the meta-analyses in our sample. As discussed above, true effect size can vary among studies included in a meta-analysis due to variation in both the samples (sample heterogeneity) and the exact design used (design heterogeneity). Heterogeneity per se cannot explain our findings, but it introduces the possibility that replications are carried out in samples or with designs not representative of the samples and designs included in the meta-analyses.

The larger the heterogeneity in a meta-analysis, the larger the scope for such replicator selection. If there is replicator selection, one would thus expect a positive correlation between heterogeneity in the meta-analyses and the meta-replication effect size difference, but no such positive correlation was observed in our data. Another indication of replicator selection would be that the effect sizes of the original studies are lower than those of the meta-analyses, as that would suggest that the original study design used in the replication is not representative of the other designs included in the meta-analysis. We find no evidence of this in our data either, which points to the original studies being reasonably representative of the wider set of designs included in the meta-analyses. However, this comparison should be interpreted cautiously due to the possibility that the original studies were associated with a larger bias, due to selective reporting and publication bias, than the other studies included in the meta-analyses. We find no evidence of our findings being explained by heterogeneity in meta-analysis and replicator selection but, at the same time, we cannot rule out that replicator selection has affected our results. Our results should therefore be interpreted cautiously, and further work on heterogeneity and replicator selection is important.

Another caveat about our results concerns the representativity of our sample. The inclusion of studies was limited by the number of preregistered multiple-laboratory replications and by studies for which we could find a matching meta-analysis. Our sample of 15 studies should thus not be viewed as being representative of meta-analysis in psychology or in other fields. In particular, the relative effect between the original studies and replication studies for the sample of studies included in our analysis is somewhat larger than that observed in previous replication projects[5,6,10]—indicating that our sample could be a select sample of studies where selective reporting is particularly prominent. In future, the number of studies using our methodology can be extended as more preregistered multiple-laboratory replications become available and as the number of meta-analyses continues to increase. We also encourage others to test our methodology for evaluation of meta-analyses on an independent sample of studies.

In a previous related study in the field of medicine, 12 large randomized, controlled trials published in four leading medical journals were compared to 19 meta-analyses published previously on the same topics[59]. They compared several clinical outcomes among the studies and found a significant difference between the meta-analyses and the large clinical trials for 12% of the comparisons. They did not provide any results for the pooled overall difference between meta-analyses and large clinical trials, but from graphical inspection of the results there does not appear to be a sizeable systematic difference. This difference in results between psychology and medicine could reflect a genuine difference between those fields, but it could also reflect the fact that even large clinical trials in medicine are subject to selective reporting or publication bias. An important difference between medicine and psychology is also the requirement of the former to register randomized controlled

trials, which may diminish the biases of published studies; no such requirement exists in psychology.

We conclude that meta-analyses produce substantially larger effect sizes than replication studies in our sample. This difference is largest for replication studies that fail to reject the null hypothesis, which is in line with recent arguments about a high false-positive rate of meta-analyses in the behavioural sciences[27,28]. Our findings suggest that meta-analysis is ineffective in fully adjusting inflated effect sizes for publication bias and selective reporting in our sample of 15 meta-analyses. We furthermore find that applying methods aiming to correct for publication bias does not substantively improve the meta-analytic results. The trim-and-fill and 3PSM bias-adjustment methods produce results similar to the conventional random effects model. PET-PEESE does adjust effect sizes downwards, but at the cost of substantial reduction in power and increase in false-negative rate. These results suggest that statistical solutions alone may be insufficient to rectify reproducibility issues in the behavioural sciences, but further research should assess whether our results for bias-adjustment methods are valid in other study samples. A potentially effective policy for reducing publication bias and selective reporting is preregistration of analysis plans before data collection, an increasing trend in psychology[60]. This has the potential to increase the credibility of both the original studies and meta-analyses, rendering the latter a more valuable tool for aggregation of research results. Future meta-analyses may thus produce effect sizes that are closer to those in replication studies.

Methods

Below we describe the data collection and conversion of effect sizes, construction of CIs in Figs. 2–6, estimation of mean effect size difference, estimation of bias-adjustment methods and the indicators used to compare the performance of bias-adjustment methods. For statistical tests based on normality, data distribution was assumed to be normal but this was not formally tested.

Data collection and conversion of effect size. First, we identified registered replications in psychology of a multiple-laboratory format. This was done by looking through issues of the journals 'Perspectives on Psychological Science' and 'Advances in Methods and Practices in Psychological Science' which, since 2014, have published studies of this format[37]. Second, we identified three Many Labs projects completed to date[7–9], each of these projects including several replication studies. In total, this yielded 62 replication effect sizes for which we searched for a corresponding meta-analysis.

For each of these 62 replications, two independent researchers searched Google Scholar for previous—published or unpublished—meta-analyses conducted on the same hypothesis as that investigated in the multiple-laboratory replication. We performed both a general search and a more restrictive search. In the former, we searched for key terms in the replication paper (for instance, 'ego depletion') in combination with 'meta analysis'. In the restricted search, we searched for 'meta analysis' and 'meta-analysis' in the database of papers in Google Scholar, citing the original study that was replicated in the multiple-laboratory replication.

We identified 39 meta-analyses deemed relevant to assessment for eligibility (see Supplementary Table 1 for details). Both researchers then assessed the eligibility of each of the meta-analyses by reading the paper alongside the replication paper and the original study paper. For unclear cases, agreement was reached through consensus and the third researcher was consulted. Twenty-one meta-analyses were excluded due to a lack of correspondence in the effects estimated in the meta-analyses and replication studies, and two were excluded due to lack of data (reducing the sample to 16 meta-analyses). The missing data related to missing information about standard errors. For meta-analyses where the standard error was not available from the published paper, we had to email the corresponding author of the meta-analysis for this criterion. In one of these, the corresponding author had died 10 years previously and we failed to find the other authors, so this meta-analysis was therefore excluded[61]. In another case, the author of the meta-analysis replied to our request but the data were missing because they had been collected more than 20 years previously[62]. Therefore, that study was also excluded from our analysis sample due to data unavailability.

For the remaining 16 meta-analyses, two were for the same replication study—the ego depletion replication[63]. Two meta-analyses concerned ego depletion (Hagger et al.[64] and Carter et al.[65]). If we had included both of these meta-analyses, they would not be independent observations and we therefore included only one, reducing our sample to 15 meta-analyses paired with 15 replication studies. In the main analysis we included the meta-analysis with the largest sample size, which was Hagger et al.[64], but we also performed a robustness test including, instead, the meta-analysis of Carter et al.[65] (in Fig. 3).

For the replication of Hauser et al.[42] there were two separate replication estimates in Klein et al.[8]. Because the meta-analysis corresponding to these two replication estimates[66] included both scenarios subjected to replication, and both replication estimates used different participants, we therefore chose to pool the replication estimates. We did this by downloading the raw data from Klein et al.[8] from the Open Science Framework and ran a random effects meta-analysis using each estimate and standard error at the laboratory level as our unit of observation, resulting in a single random effects meta-analytical estimate for this replication.

For the replication of Schooler and Engstler-Schooler[47] by Alogna et al.[67] there were two replication studies, one of study 1 from Schooler and Engstler-Schooler and one of study 4 from Schooler and Engstler-Schooler. As both of these replication studies were included in the meta-analysis by Meissner and Brigham[68], we pooled these replication studies by conducting a random effects meta-analysis using the effect size and standard error for all primary studies entering the replication study. Our reason for treating each estimates as statistically independent is that different subjects were used in the replication of studies 1 and 4.

For the meta-analysis of Kivikangas et al.[69] corresponding to Graham, Haidt and Nosek[44], there were three separate estimates of the three 'binding foundations' in the meta-analysis, whereas the replication presents a single pooled measure. To ensure that the observations included in our overall measure were statistically independent, we followed our criterion of selecting the most precise estimate for inclusion in the main results, and therefore included the 'authority' estimate from Kivikangas et al.[69]. For robustness, in Fig. 3 we present the results using the two other possible choices of meta-analytic estimate instead.

For the 15 meta-analyses comprising the final sample, 11 included the original study which was replicated in the corresponding replication study. This suggests that, for these 11 studies, the meta-analysts and the authors of this article made the same decision regarding estimation by the meta-analyses and replication studies of the same effects. For the remaining four meta-analyses, we made the decision that these study the same effects as in the corresponding replication studies. We comment on these four cases below:

- Srull and Wyer[51]: The meta-analysts[70] explicitly state that they intended to include this study, but suspected a statistical error and therefore chose not to include it (thus, they clearly viewed the design of the original study as belonging in the meta-analysis).
- Sripada et al.[48]: The meta-analysis by Hagger et al.[64] was published before the original study by Sripada et al.[48], which was replicated in 2015. Moreover, the e-letter task included in Sripada et al.[48] is an electronic modification of a corresponding design used in Baumeister et al.[71], which is included in the meta-analysis of Hagger et al.[64]. For the subsequent meta-analysis of Carter et al.[6], which is included in the robustness test rather than the meta-analysis of Hagger et al.[64], we found no obvious explanation for not including Sripada et al.[48], but Carter et al.[65] is a widely published study and argues that it constitutes an improvement over Hagger et al.[64].
- Oppenheimer et al.[38]: The meta-analysis by Roth et al.[72] does not include or cite the original study by Oppenheimer et al.[38], but it cites and includes the study by Thaler.[73]. Because the original study by Oppenheimer et al.[38] adapts the sunk cost question directly from the study by Thaler[73] and uses the same wording of the question, we find no obvious explanation for why the Oppenheimer et al.[38] study was not included in the meta-analysis.
- Graham, Haidt and Nosek[44]: The meta-analysis by Kivikangas et al.[69] does not include the original study but cites it as seminal for the literature reviewed, and the authors in fact use the original study as the starting point for defining the period used to search for relevant studies. The reason for not including the original study is not stated in the meta-analysis.

After deciding on an analysis sample, we obtained the relevant data from the replication study either from the information in the published paper or from datasets publicly available on the Open Science Framework. For the meta-analyses, as far as possible we obtained data on the summary effect and the standard error of the summary effect from the information available in the published paper (if the standard error was not directly reported, we derived it from the 95% CI of the standardized Cohen's d effect size). In all but two of the included meta-analyses, it was possible to extract the relevant information directly from the paper; in those two cases we had to email the corresponding author of the meta-analysis.

To compare results across studies we needed to ensure that the results were measured using the same effect size metric. In most cases the effect size reported is a standardized Cohen's d measure. However, in a minority of the cases a correlation coefficient (Pearson's r or Fisher's z) is reported, and some studies report an unconverted 'natural unit' of the effect as their main measure—for example, the percentage point difference between treatments. To put all our results on the same scale we converted all effects to Cohen's d, with the exception of one meta-replication pair where effect sizes as noted above were measured using Cohen's q (but where we performed a robustness test without this meta-replication pair). Two of the meta-analyses also measured effect sizes in Hedges' g units (see Supplementary Table 3), but because that is very similar to Cohen's d we did not convert the effect sizes of these meta-analyses.

In cases where a correlation coefficient is reported, we convert the effect sizes to a Cohen's d measure using the following formula:

$$d = \frac{2r}{\sqrt{1-r^2}}$$

and Fisher's z is converted to r according to the inverse of the Fisher transformation, so that

$$r = \frac{\exp(2z)-1}{\exp(2z)+1}$$

which is again converted to d using the above formula. These transformations follow from statistical theory and rely on an assumption that the data follow a bivariate normal distribution[74]. In cases where the main effect size is reported as the difference in percentage points between conditions, we divide the estimated treatment effect by the standard deviation of the dependent variable and obtained a Cohen's d measure of the effect size.

It should also be noted that we initially identified 17 meta-replication pairs, but we discovered that two of these were not actually matches after obtaining data on the individual studies included in each of the 17 meta-analyses (needed for the estimation of the meta-analysis bias-adjustment methods, see below). These two meta-analyses did include one study from the original paper[75,76] replicated in the multiple-laboratory replication project, but it was not the same study as that replicated (these two original studies reported the results of several studies; one was replicated in the replication study and the other was included in the meta-analysis). This was not evident from the published meta-analyses, but could be seen only after obtaining data on the individual studies included in each of the meta-analyses.

Construction of CIs. The point estimates for the effect sizes in Fig. 2a are the mean effect sizes (converted to Cohen's d) reported in the meta-analyses studies and replication studies; the 95% CIs are constructed as $\pm 1.96 \times$ s.e.m and the 99.5% CIs are constructed as $\pm 2.807 \times$ s.e.m. (based on the reported s.e.m. in both meta-analyses and replication studies). The point estimates of the effect size differences in Fig. 2b denote the difference between meta-analytic and replication effect size in Fig. 2a and, as above, the 95% CIs are constructed as $\pm 1.96 \times$ s.e.m. and the 99.5% CIs are constructed as $\pm 2.807 \times$ s.e.m. (the standard error of the difference in meta-analytic and replication effect size as estimated by z-test). The CIs in Figs. 3–6 are constructed in the same way.

Estimation of mean effect size difference. We used random effects meta-analysis to estimate mean effect size difference across the 15 meta-replication pairs. The random effects model treats the true parameters as different draws from an overall distribution and weights each study by the inverse of the sum of within- and between-study variance[74]. Statistical inference in the random effects model is built on the assumption that random effects are normally distributed[77].

For replications that are part of the Many Labs project, the same individuals participated in several replication studies within each project. In our study, three replications from Many Labs 1 are based on the same individuals and two from the Many Labs 3 project are based on the same individuals. In Many Labs 2, data were collected in two slates with different samples, and two of the replications are based on individuals in slate 1—one from Many Labs 2 is based on individuals from slates 1 and 2, and another is based on individuals from slate 2. This introduces a violation of independence among some of the replication studies, which could affect the estimated standard errors of the pooled effect size differences. In the sub-group analysis in Fig. 3, we report results separately for replication studies from the Many Labs projects and replications from the registered report replication projects (where there is no violation of independence among the replication studies). The results in both sub-groups are similar.

Estimation of meta-analysis bias-adjustment methods. In Figs. 5 and 6, we estimate three different bias-correction models for our 15 meta-analyses and compare these to both the non-adjusted random effects results and the replication effect sizes. To be able to do these estimations we needed the results of the individual studies included in each of the 15 meta-analyses (mean effect size and standard error). For some meta-analyses these data had been posted by the authors, and for the remainder we obtained them after emailing the authors. The sections below detail the general features of each bias-correction method and how we implemented them in our analysis.

Trim-and-fill. Trim-and-fill, developed by Duval and Tweedie[11], is an algorithm that aims to identify 'missing values' from a distribution of studies in a standard meta-analysis, imputes these missing values into the data and then computes the selection-corrected effect by conducting a meta-analysis on the full dataset including both the original studies and the imputed values. Formally, the trim-and-fill method estimates a conventional meta-analytic average (either fixed or random effects) for the n studies in the dataset but assumes that there are a number of missing studies, denoted by k_0. The iterative trim-and-fill algorithm proceeds by estimating the number of missing studies using one of several possible estimators, then computes an estimate for the missing studies and finally computes a selection-corrected weighted average of both originally included and imputed studies. We use the random effects model as the baseline model when implementing the trim-and-fill algorithm, because this is the conventional model choice for meta-analysts.

PET-PEESE. PET-PEESE is a regression-based approach suggested by Stanley and Doucouliagos[31]. The basic idea is to run a meta-regression of the effect size on the standard error and take the constant term as the measure of the true effect free from selection bias. The PET runs the meta-regression

$$Y_i = \gamma_0 + \alpha \mathrm{SE}_i + \varepsilon_i,$$

where i indexes the primary study, SE is the study-level standard error and ε_i is an idiosyncratic error term. The meta-regression uses $\frac{1}{\mathrm{SE}_i^2}$ as weights. The estimate of γ_0 is treated as a measure of the selection-corrected true effect (the effect that would result in a setting with no sampling variation). PEESE instead replaces SE in the above regression equation by the squared standard error. PET-PEESE, which we employ in our estimations, is a conditional estimator that uses the PET-estimate if PET fails to reject the null hypothesis that $\gamma_0 = 0$, and uses the PEESE estimate if PET rejects the null hypothesis. We use PET-PEESE in our analysis as suggested by Stanley and Doucouliagos[31]. We use PET if PET fails to reject the null hypothesis that $\gamma_0 = 0$ at the 5% level, and otherwise we use PEESE.

3PSM. The 3PSM method was developed by Hedges and Vevea[29] and is a sophisticated selection model that is estimated through maximum likelihood. Similar to conventional meta-analysis, the model posits that effect sizes are distributed as $\hat{\theta}_i \sim N[\theta_i, \sigma_i^2]$ and that true effects are distributed as $\theta_i \sim N[\mu, \tau^2]$. The model allows for selective reporting by representing the likelihood that a P value of p_i is observed by a (step) weight function $w{:}p_i \rightarrow \omega_i$ (where ω_i is the relative weight assigned to a given P value interval), and is solved by maximizing the joint log-likelihood function for the data with respect to the parameters θ, τ^2 and ω using the Newton–Raphson algorithm.

Our implementation of 3PSM follows the default choice for the weight function, with a cut-off of one-tailed $P = 0.025$, so that we allow for different weights for observations above and below this threshold. This specification of the weight function is the same as that chosen in the simulations of Carter et al.[34], who found that this implementation of 3PSM tends to perform very well under a wide set of conditions. We implemented 3PSM by uploading our data to the online app. programmed by Vevea and Coburn, available at https://vevealab.shinyapps.io/WeightFunctionModel/. For the study by Rabelo et al.[78], because there was an insufficient number of studies to implement the model, we used the closest possible alternative threshold $P < 0.025$ that returned an estimate. The model returned an estimate for a cut-off of $P = 0.024$, so that was used for this particular study.

Indicators used to compare performance of the bias-adjustment methods. We used a number of indicators to compare the performance of the three bias-adjustment methods and the random effects model to the replication studies. The results of these indicators are reported in Table 1. We include the following indicators:

- False-positive rate: The starting point here is those studies where the replication study cannot reject the null hypothesis, and the indicator is measured as the fraction of these studies where the meta-analysis finds a significant positive effect size. Note that this measure assumes that the null hypothesis is true for all studies where the replication study cannot reject the null hypothesis. We report this indicator for both the 0.5 and 5% significance level.
- False-negative rate: The starting point here is those studies where the replication study finds a significant positive effect size, and the indicator is measured as the fraction of these studies where the meta-analysis cannot reject the null hypothesis. Note that this measure assumes that the hypothesis is true if the replication study finds a significant positive effect. We report this indicator for both the 0.5 and 5% significance level. One replication study[8] found a significant negative effect size (see Fig. 2) and is thus not included among the studies used to estimate either the false-negative or -positive rate.
- Mean meta-replication difference, random effects: This is a random effects meta-analysis estimate of the mean effect size differences across the 15 meta-replication pairs. We report the random effects mean and the z-statistic of this mean (the mean divided by standard error).
- Mean meta-replication difference, unweighted. This is the mean difference between the meta-analysis and the replication study for the 15 meta-replication pairs; this measure is referred to as the mean error by Carter et al.[34]. It is a measure of by how much meta-analysis overestimates the replication effect sizes.
- Overestimation factor: This is estimated as the unweighted mean meta-analytic effect size divided by the unweighted mean replication study effect size. It is a measure of the relative overestimation of effect size in meta-analysis.
- Root mean squared error: This is a measure of the prediction error of the meta-analysis, and can be seen as a measure of the precision of the meta-analysis results.
- Mean MDE: This denotes the minimally detectable effect size at 80% power, which is the effect size that the meta-analysis has 80% power to detect. We

take the mean of the MDE across the 15 meta-analyses, which is a measure of their average power. We report this indicator for both the 0.5 and 5% significance level. To estimate the MDE for each meta-analysis we use the z-distribution, where it is computed as 3.65 × s.e.m. for the 0.5% significance level and 2.8 × s.e.m. for the 5% significance level.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data used in this paper are posted at the project's OSF repository (link: https://osf.io/vw3p6).

Code availability

The analysis code for all analyses are available at the project's OSF repository (link: https://osf.io/vw3p6).

Received: 1 March 2019; Accepted: 13 November 2019;
Published online: 23 December 2019

References

1. Siddaway, A. P., Wood, A. M. & Hedges, L. V. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annu. Rev. Psychol.* **70**, 747–770 (2019).
2. Cumming, G. The new statistics: why and how. *Psychol. Sci.* **25**, 7–29 (2014).
3. Stanley, T. D. Wheat from chaff: meta-analysis as quantitative literature review. *J. Econ. Perspect.* **15**, 131–150 (2001).
4. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182 (2018).
5. Camerer, C. F. et al. Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
6. Camerer, C. F. et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637 (2018).
7. Klein, R. A. et al. Investigating variation in replicability: a "Many Labs" replication project. *Soc. Psychol.* **45**, 142–152 (2014).
8. Klein, R. A. et al. Many Labs 2: investigating variation in replicability across samples and settings. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2018).
9. Ebersole, C. R. et al. Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82 (2016).
10. Estimating, O. S. C. The reproducibility of psychological science. *Science* **349**, aac4716 (2015).
11. Duval, S. & Tweedie, R. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *J. Am. Stat. Assoc.* **95**, 89–98 (2000).
12. Ioannidis, J. P. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
13. Ioannidis, J. P. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).
14. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
15. Gelman, A. & Carlin, J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspect. Psychol. Sci.* **9**, 641–651 (2014).
16. Gelman, A. & Loken, E. The statistical crisis in science. *Am. Sci.* **102**, 460 (2014).
17. Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. Star wars: the empirics strike back. *Am. Econ. J. Appl. Econ.* **8**, 1–32 (2016).
18. Andrews, I. & Kasy, M. Identification of and correction for publication bias. *Am. Econ. Rev.* **109**, 2766–2794 (2019).
19. Schäfer, T. & Schwarz, M. A. The meaningfulness of effect sizes in psychological research: differences between sub-disciplines and the impact of potential biases. *Front. Psychol.* **10**, article 813 (2019).
20. John, L. K., Loewenstein, G. & Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23**, 524–532 (2012).
21. Franco, A., Malhotra, N. & Simonovits, G. Publication bias in the social sciences: unlocking the file drawer. *Science* **345**, 1502–1505 (2014).
22. Franco, A., Malhotra, N. & Simonovits, G. Underreporting in political science survey experiments: comparing questionnaires to published results. *Polit. Anal.* **23**, 306–312 (2015).
23. Sterne, J. A., Gavaghan, D. & Egger, M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J. Clin. Epidemiol.* **53**, 1119–1129 (2000).
24. Rothstein, H. R., Sutton, A. J. & Borenstein, M. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (Wiley, 2005).
25. Schwarzer, G., Carpenter, J. R. & Rücker, G. in *Meta-analysis with R. Use R!* 107–141 (Springer, 2015).
26. Polanin, J. R., Tanner-Smith, E. E. & Hennessy, E. A. Estimating the difference between published and unpublished effect sizes: a meta-review. *Rev. Educ. Res.* **86**, 207–236 (2016).
27. Nelson, L. D., Simmons, J. & Simonsohn, U. Psychology's renaissance. *Annu. Rev. Psychol.* **69**, 511–534 (2018).
28. Vosgerau, J., Simonsohn, U., Nelson, L. D. & Simmons, J. P. 99% impossible: a valid, or falsifiable, internal meta-analysis. *J. Exp. Psychol. Gen.* **148**, 1628 (2019).
29. Vevea, J. L. & Hedges, L. V. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* **60**, 419–435 (1995).
30. Hedges, L. V. Modeling publication selection effects in meta-analysis. *Stat. Sci.* **7**, 246–255 (1992).
31. Stanley, T. D. & Doucouliagos, H. Meta-regression approximations to reduce publication selection bias. *Res. Synth. Methods* **5**, 60–78 (2014).
32. Iyengar, S. & Greenhouse, J. B. Selection models and the file drawer problem. *Stat. Sci.* **3**, 109–117 (1988).
33. Simonsohn, U., Nelson, L. D. & Simmons, J. P. *P*-curve and effect size: correcting for publication bias using only significant results. *Perspect. Psychol. Sci.* **9**, 666–681 (2014).
34. Carter, E. C., Schönbrodt, F. D., Gervais, W. M. & Hilgard, J. Correcting for bias in psychology: a comparison of meta-analytic methods. *Adv. Methods Pract. Psychol. Sci.* **2**, 115–144 (2019).
35. McShane, B. B., Böckenholt, U. & Hansen, K. T. Adjusting for publication bias in meta-analysis: an evaluation of selection methods and some cautionary notes. *Perspect. Psychol. Sci.* **11**, 730–749 (2016).
36. Stanley, T. D. Limitations of PET-PEESE and other meta-analysis methods. *Soc. Psychol. Personal. Sci.* **8**, 581–591 (2017).
37. Simons, D. J., Holcombe, A. O. & Spellman, B. A. An introduction to registered replication reports at perspectives on psychological science. *Perspect. Psychol. Sci.* **9**, 552–555 (2014).
38. Oppenheimer, D. M., Meyvis, T. & Davidenko, N. Instructional manipulation checks: detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* **45**, 867–872 (2009).
39. Tversky, A. & Kahneman, D. The framing of decisions and the psychology of choice. *Science* **211**, 453–458 (1981).
40. Husnu, S. & Crisp, R. J. Elaboration enhances the imagined contact effect. *J. Exp. Soc. Psychol.* **46**, 943–950 (2010).
41. Schwarz, N., Strack, F. & Mai, H.-P. Assimilation and contrast effects in part-whole question sequences: a conversational logic analysis. *Public Opin. Q.* **55**, 3–23 (1991).
42. Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R. & Mikhail, J. A dissociation between moral judgments and justifications. *Mind Lang.* **22**, 1–21 (2007).
43. Critcher, C. R. & Gilovich, T. Incidental environmental anchors. *J. Behav. Decis. Mak.* **21**, 241–251 (2008).
44. Graham, J., Haidt, J. & Nosek, B. A. Liberals and conservatives rely on different sets of moral foundations. *J. Personal. Soc. Psychol.* **96**, 1029 (2009).
45. Jostmann, N. B., Lakens, D. & Schubert, T. W. Weight as an embodiment of importance. *Psychol. Sci.* **20**, 1169–1174 (2009).
46. Monin, B. & Miller, D. T. Moral credentials and the expression of prejudice. *J. Personal. Soc. Psychol.* **81**, 33 (2001).
47. Schooler, J. W. & Engstler-Schooler, T. Y. Verbal overshadowing of visual memories: some things are better left unsaid. *Cogn. Psychol.* **22**, 36–71 (1990).
48. Sripada, C., Kessler, D. & Jonides, J. Methylphenidate blocks effort-induced depletion of regulatory control in healthy volunteers. *Psychol. Sci.* **25**, 1227–1234 (2014).
49. Rand, D. G., Greene, J. D. & Nowak, M. A. Spontaneous giving and calculated greed. *Nature* **489**, 427 (2012).
50. Strack, F., Martin, L. L. & Stepper, S. Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *J. Personal. Soc. Psychol.* **54**, 768 (1988).
51. Srull, T. K. & Wyer, R. S. The role of category accessibility in the interpretation of information about persons: some determinants and implications. *J. Personal. Soc. Psychol.* **37**, 1660 (1979).
52. Mazar, N., Amir, O. & Ariely, D. The dishonesty of honest people: a theory of self-concept maintenance. *J. Mark. Res.* **45**, 633–644 (2008).
53. Benjamin, D. J. et al. Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6 (2018).
54. Fanelli, D., Costas, R. & Ioannidis, J. P. Meta-assessment of bias in science. *Proc. Natl Acad. Sci. USA* **114**, 3714–3719 (2017).
55. Augusteijn, H. E., van Aert, R. & van Assen, M. A. The effect of publication bias on the Q test and assessment of heterogeneity. *Psychol. Methods* **24**, 116 (2019).
56. Stanley, T., Carter, E. C. & Doucouliagos, H. What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull.* **144**, 1325–1346 (2018).

57. van Aert, R. C., Wicherts, J. M. & van Assen, M. A. Conducting meta-analyses based on *P* values: reservations and recommendations for applying *P*-uniform and *P*-curve. *Perspect. Psychol. Sci.* **11**, 713–729 (2016).
58. Simonsohn, U., Nelson, L. D. & Simmons, J. P. *P*-curve: a key to the file-drawer. *J. Exp. Psychol. Gen.* **143**, 534 (2014).
59. LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J. & Derderian, F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N. Engl. J. Med.* **337**, 536–542 (1997).
60. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606 (2018).
61. Mullen, B. Strength and immediacy of sources: a meta-analytic evaluation of the forgotten elements of social impact theory. *J. Personal. Soc. Psychol.* **48**, 1458 (1985).
62. Holleman, B. Wording effects in survey research using meta-analysis to explain the forbid/allow asymmetry. *J. Quant. Linguist.* **6**, 29–40 (1999).
63. Hagger, M. S. et al. A multilab preregistered replication of the ego-depletion effect. *Perspect. Psychol. Sci.* **11**, 546–573 (2016).
64. Hagger, M. S., Wood, C., Stiff, C. & Chatzisarantis, N. L. Ego depletion and the strength model of self-control: a meta-analysis. *Psychol. Bull.* **136**, 495 (2010).
65. Carter, E. C., Kofler, L. M., Forster, D. E. & McCullough, M. E. A series of meta-analytic tests of the depletion effect: self-control does not seem to rely on a limited resource. *J. Exp. Psychol. Gen.* **144**, 796 (2015).
66. Feltz, A. & May, J. The means/side-effect distinction in moral cognition: a meta-analysis. *Cognition* **166**, 314–327 (2017).
67. Alogna, V. et al. Registered replication report: Schooler and Engstler-Schooler (1990). *Perspect. Psychol. Sci.* **9**, 556–578 (2014).
68. Meissner, C. A. & Brigham, J. C. A meta-analysis of the verbal overshadowing effect in face identification. *Appl. Cogn. Psychol.* **15**, 603–616 (2001).
69. Kivikangas, J. M., Lönnqvist, J.-E. & Ravaja, N. Relationships between moral foundations and political orientation–local study and meta-analysis. in *Annual Convention of Society for Personality and Social Psychology* https://doi.org/10.13140/RG.2.1.2277.0964 (2016).
70. DeCoster, J. & Claypool, H. M. A meta-analysis of priming effects on impression formation supporting a general model of informational biases. *Personal. Soc. Psychol. Rev.* **8**, 2–27 (2004).
71. Baumeister, R. F., Bratslavsky, E. & Muraven, M. Ego depletion: is the active self a limited resource? *J. Personal. Soc. Psychol.* **74**, 1252–1265 (2018).
72. Roth, S., Robbert, T. & Straus, L. On the sunk-cost effect in economic decision-making: a meta-analytic review. *Bus. Res.* **8**, 99–138 (2015).
73. Thaler, R. Mental accounting and consumer choice. *Mark. Sci.* **4**, 199–214 (1985).
74. Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. *Introduction to Meta-analysis* (Wiley, 2011).
75. Galinsky, A. D., Magee, J. C., Inesi, M. E. & Gruenfeld, D. H. Power and perspectives not taken. *Psychol. Sci.* **17**, 1068–1074 (2006).
76. Finkel, E. J., Rusbult, C. E., Kumashiro, M. & Hannon, P. A. Dealing with betrayal in close relationships: does commitment promote forgiveness? *J. Personal. Soc. Psychol.* **82**, 956 (2002).
77. Higgins, J. P., Thompson, S. G. & Spiegelhalter, D. J. A re-evaluation of random-effects meta-analysis. *J. R. Stat. Soc. Ser. A* **172**, 137–159 (2009).
78. Rabelo, A. L., Keller, V. N., Pilati, R. & Wicherts, J. M. No effect of weight on judgments of importance in the moral domain and evidence of publication bias from a meta-analysis. *PloS One* **10**, e0134808 (2015).

Acknowledgements

For financial support we thank J. Wallander and the Tom Hedelius Foundation (grant no. P2015-0001:1), the Swedish Foundation for Humanities and Social Sciences (grant no. NHS14-1719:1) and the Meltzer Fund in Bergen. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

A.K., E.S. and M.J. designed research and wrote the paper. A.K. and E.S. collected and analysed data. All authors approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0787-z.

Correspondence and requests for materials should be addressed to M.J.

Peer review information Primary handling editor: Aisha Bradshaw

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Magnus Johannesson

Last updated by author(s): Oct 8, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All data was manually collected (no software was used)
Data analysis	STATA 16

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data and code used in the paper is available at the project's OSF repository (https://osf.io/vw3p6) and will be made public upon publication.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study is a "meta-meta analysis" of the psychological literature
Research sample	Sample of meta-analyses (published and unpublished) and published replication studies
Sampling strategy	We started with a sample of all pre-registered multiple lab replication studies, and our finally sample features pairs of studies for which there was at least one meta-analysis studying the same research question as the one investigated in the replication study
Data collection	Researchers collected data by searching through Google Scholar, and in some cases by e-mailing the original authors for the data
Timing	We started screening for replication studies and matching meta-analyses in March 2018 and completed this screening process in September 2018. We then collected data for the identified studies until June 2019.
Data exclusions	No data that matched our inclusion criteria were excluded
Non-participation	N/A
Randomization	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging