US studies may overestimate effect sizes in softer research

Daniele Fanelli[a,1] and John P. A. Ioannidis[b,c,d]

[a]Science, Technology and Innovation Studies, The University of Edinburgh, Edinburgh EH1 1LZ, United Kingdom; and [b]Stanford Prevention Research Center, Department of Medicine, and Departments of [c]Health Research and Policy and [d]Statistics, School of Humanities and Sciences, Stanford University, Stanford, CA 94305

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved July 19, 2013 (received for review February 14, 2013)

Many biases affect scientific research, causing a waste of resources, posing a threat to human health, and hampering scientific progress. These problems are hypothesized to be worsened by lack of consensus on theories and methods, by selective publication processes, and by career systems too heavily oriented toward productivity, such as those adopted in the United States (US). Here, we extracted 1,174 primary outcomes appearing in 82 meta-analyses published in health-related biological and behavioral research sampled from the Web of Science categories Genetics & Heredity and Psychiatry and measured how individual results deviated from the overall summary effect size within their respective meta-analysis. We found that primary studies whose outcome included behavioral parameters were generally more likely to report extreme effects, and those with a corresponding author based in the US were more likely to deviate in the direction predicted by their experimental hypotheses, particularly when their outcome did not include additional biological parameters. Nonbehavioral studies showed no such "US effect" and were subject mainly to sampling variance and small-study effects, which were stronger for non-US countries. Although this latter finding could be interpreted as a publication bias against non-US authors, the US effect observed in behavioral research is unlikely to be generated by editorial biases. Behavioral studies have lower methodological consensus and higher noise, making US researchers potentially more likely to express an underlying propensity to report strong and significant findings.

publish or perish | soft science | research bias | questionable research practices | scientific misconduct

Science is a struggle for truth against methodological, psychological, and sociological obstacles. Increasing efforts are devoted to studying unconscious and conscious biases in research and publication, because these represent a threat for human health, economic resources, and scientific progress (1). Thanks to refinements of meta-analytical techniques, a variety of bias-related patterns have been identified. Thus, the old notion of a "file-drawer" effect, in which statistically significant results are more likely to be published (2), is now integrated with evidence of a decline effect, an early-extremes effect (Proteus phenomenon), and numerous other reporting biases (1, 3–9). Other methodological problems, including expectancy effects and scientific misconduct, have been noted for decades or even centuries. However, scholars only recently started to assess systematically their prevalence across fields and countries, study their causes, and openly discuss general solutions (e.g., 10–13).

The publication of false, exaggerated, and falsified findings is believed to be more common in research fields where replication is difficult, theories are less clear, and methods are less standardized, because researchers have more "degrees of freedom" to produce the results they expect (6, 14). Behavioral methodologies, in particular, have been considered at higher risk of bias at least since the 1970s (15, 16). The intuitive assumption that theoretical and methodological "softness" might increase the prevalence of expectation biases is supported by direct studies of the literature, which suggest that the proportion of papers reporting "positive" outcomes increases moving from the physical to the medical and social sciences and, independent of discipline, is higher among social and behavioral studies on people, compared with nonbehavioral studies and studies on nonhuman behavior (17).

Biases in research and reporting are presumably exacerbated by systems of publication and career evaluation that reward impact and productivity over quality and replicability. Many concerns have been expressed, in particular, for the "publish-or-perish" philosophy that has long characterized research in the United States and is increasingly taken up in other countries (e.g., 18–21). Such concerns are increasingly supported by evidence. Researchers working the United States report, in surveys, higher pressures than those in most other countries (22). At least two independent meta-analyses, one in economics and one in genetic association studies, had noted signs of a larger publication bias among papers from the United States (or North America) (23, 24). The proportion of reported positive results has increased in recent years in most social and biomedical sciences and is greater in US studies, particularly among the most academically productive states (25, 26).

In sum, past observations would suggest that studies using putatively softer methodologies and studies from the United States might tend to overestimate findings. Intrinsic limitations, however, make this evidence inconclusive. Studies that assessed the frequency of positive results in random samples of papers cannot separate the effects of research and reporting biases from the effects of editorial decisions and methodological superiority (17, 26). Meta-analyses that noted a higher publication bias from the United States, on the other hand, are limited in size and were not originally intended to test that effect—conditions that increase the likelihood of these observations' being simply false positives.

To obtain additional evidence on these controversial issues, here we analyzed 1,174 outcomes reported in 82 meta-analyses published between 2009 and 2012 in journals classified by the Web of Science in the subject categories of Genetics & Heredity (GH) and Psychiatry (PS). We chose these two categories for three main reasons. First, most journals in these (health-related) areas follow established practices of presenting results in forest plots, which ensured that primary data would be readily available from the meta-analyses we sampled. Second, much previous evidence on publication and related biases had been produced in these research areas (e.g., 1, 3, 4, 12, 17, 27–32). Third, these two subject categories covered a spectrum of methodological softness, measuring

Author contributions: D.F. conceived the project; D.F. and J.P.A.I. designed research; D.F. performed research; D.F. and J.P.A.I analyzed the data; and D.F. and J.P.A.I. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: dfanelli@exseed.ed.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1302997110/-/DCSupplemental.

parameters that are purely biological [nonbehavioral (NB)], purely behavioral (BE), or a combination of the two [biobehavioral (BB)].

Results

We first divided each primary effect size by the summary effect size of its meta-analysis. As shown in Fig. 1 (Fig. S1 shows the complete range), primary studies from the biological (NB) meta-analyses are characterized by a relatively ordered distribution, with studies deviating from the "true" effect (summary effect size) in inverse proportion to their precision (i.e., by the inverse of their SE). This is exactly what we expect when fluctuations in study outcomes are relatively unbiased and determined primarily by sampling variance. Studies from the behavioral meta-analyses, however, had visibly less-ordered distributions, their dispersion being greater and less dependent on study size (Fig. 1). This is particularly the case for those having purely behavioral outcomes (BE) as opposed to those combining behavioral and nonbehavioral measurements (BB).

We tested for statistical differences in the distributions showed in Fig. 1 using two measures, which we called "deviation score" and "expectation factor." The deviation score simply folds the distributions in Fig. 1 and thus measures the absolute tendency to deviate from the summary effect size. The expectation factor consists of a dummy variable that gives a score of 1 to meta-analyses whose experimental hypothesis predicted odds ratio (OR) >1 and a score of 0 to those predicting the opposite. A significant positive effect of this predictor suggests that values of the deviation score are not distributed at random because they favor the experimental hypothesis (*Methods* gives details).

To account for the possible nonindependence of values reported within each meta-analysis [for example, because in some fields extreme values are more likely to be published (4)], we included a random intercept in regression models (33, 34, 35). Indeed, random effects at the meta-analysis level alone accounted for circa 14% of the variance in deviation scores (variance of intercept ± SD = 0.0043 ± 0.655, residual = 0.0272 ± 0.1649). Controlling for this effect, study size, measured by the study's SE, was the strongest predictor of deviation score (0.24[0.21, 0.28]), as we would expect because small studies' outcomes are more likely to fluctuate. Henceforth, all generalized linear model results are weighted by SE, as required in metaregression, unless differently specified. The chronological order of appearance of a study within the meta-analysis had a small, yet statistically significant, additional effect (inverse-variance weighted multilevel regression: 0.01[0.01, 0.02]).

Controlling for the effects above, BE/BB studies had significantly larger deviation scores than NB (0.07[0.06, 0.1]), the effect being stronger in BE than in BB (respectively, 0.08[0.06, 0.11] and 0.05[0.03, 0.09]). Geographical location of corresponding author modulated these effects. The likelihood to deviate was significantly higher for US studies compared with those from all countries (Table 1), and particularly European Union-15 countries (Table 2), but the effect only occurred in BE/BB, and was stronger in BE. The effect of study size on deviation, instead, was weaker in BB/BE than in NB and was stronger for non-US countries, particularly for European Union-15 countries and in BE (Tables 1 and 2).

Interaction terms between country and the expectation factor—which could only be applied to BE/BB (*Methods*)—had significantly stronger values for the United States (Tables 3 and 4), which suggests that the United States are more likely than other countries to overestimate effects in the direction that favors the experimental hypothesis.

We assessed a number of confounding factors and alternative approaches to measuring and analyzing these patterns, finding results in most cases to be robust. We hypothesized that converting some of the original outcomes (i.e., Cohen's d and Hedges' g) to odds ratio might have introduced a bias (*Methods* gives further details), and that meta-analyses from non-US and non-European countries might be subject to methodological shortcomings, as suggested by previous studies (29). Although we did detect small significant effects linked to both these factors, the central findings of this study were independent of these (indeed, the effects were stronger once controlling for these factors) as well as of year of publication of meta-analysis and number of authors of primary study, which we tested as an additional proxy of study size and quality (Tables S1 and S2). Similar results were obtained if the deviation score was calculated using a fixed-effects model or using the summary estimate reported in the original meta-analysis (Tables S3 and S4). We repeated the analyses using a different measure of deviation (i.e., scaling primary effect sizes around meta-analytical summary estimates), again observing similar patterns (e.g., Table S5). When analyses were run on an unweighted scaled deviation, however, we obtained extreme contradictory values between disciplines, and no overall US effect (e.g., Table S6). Finally, we tried rescaling the SEs to cancel differences in average study size between meta-analyses. Despite the loss of information, effects were unchanged in direction and magnitude, although only those for the deviation score were nominally statistically significant (Tables S7 and S8).

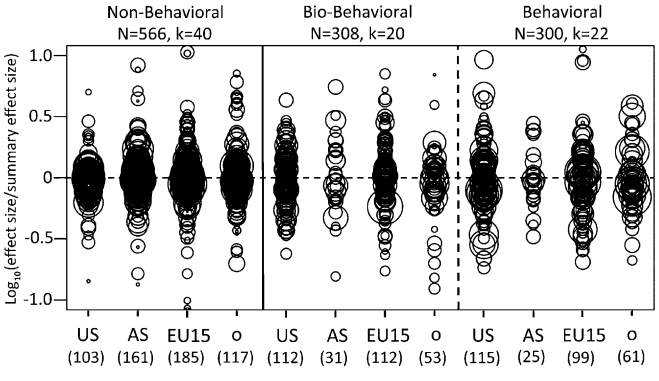

Fig. 1. Magnitude of effect sizes of primary studies relative to the summary effect size in their respective meta-analysis, partitioned by geographical origin of their corresponding author and by type of study. Size of circle is proportional to study size, measured by the SE. For illustration purposes, here size is equal to ln(2/SE). The value of 0 corresponds to a perfect matching between primary study and the summary effect size calculated from the study's meta-analysis, using a random effects model. The range of values was limited to −1 +1 to show more details. The complete range is given in Fig. S1. AS, China, Japan, South Korea, Taiwan, Singapore, and India; EU15, European Union-15 countries; o, all other countries; US, United States.

Discussion

We sampled 82 recent meta-analyses, extracted nearly 1,200 primary study outcomes, and measured how each of these latter deviated from the overall summary estimate—which by assumption should approximate the true effect that primary studies were trying to measure. Nonbehavioral biological studies largely drawn from the genetics literature fluctuated mostly because of sampling error and showed greater small-study effects, particularly when corresponding authors were not based in the United States. Conversely, behavioral studies were significantly more likely to report extreme effects, and those with a corresponding author in the US were significantly more likely to deviate in the direction predicted by their experimental hypotheses, particularly when their outcome did not include biological (e.g., physiological) parameters. These findings are best explained as the effect of an interaction between the strength of researchers' expectancy effects and their field's level of softness (i.e., low methodological consensus and high complexity of subject matter).

Table 1. Predictors of over/underestimation, United States vs. all other countries

Predictor	Nonbehavioral ($k = 40$, $n = 566$)	Behavioral, all ($k = 42$, $n = 608$)	Biobehavioral ($k = 20$, $n = 308$)	Behavioral ($k = 22$, $n = 300$)
(Intercept)	0.42 [0.40, 0.46]	0.55 [0.51, 0.56]	0.51 [0.47, 0.54]	0.57 [0.50, 0.59]
United States vs. rest	−0.02 [−0.06, 0.00]	0.03 [0.02, 0.06]	0.03 [0.00, 0.07]	0.04 [0.01, 0.07]
Study size (SE)	0.43 [0.27, 0.53]	0.11 [0.07, 0.23]	0.20 [0.11, 0.31]	0.06 [0.01, 0.29]
Pub. order	0.02 [0.00, 0.03]	0.00 [−0.01, 0.01]	0.01 [0.00, 0.05]	0.00 [−0.02, 0.01]
USA*SE	−0.21 [−0.47, 0.22]	−0.19 [−0.31, −0.03]	−0.16 [−0.34, 0.12]	−0.22 [−0.46, −0.02]
USA*pub. order	−0.02 [−0.05, 0.01]	0.00 [−0.02, 0.03]	−0.02 [−0.06, 0.01]	0.01 [−0.02, 0.05]

Likelihood of a primary study within a meta-analysis to deviate from the summary effect size (deviation score, i.e., double square root-transformed absolute value of deviation values in Fig. 1), depending on its size (SE), its chronological order of appearance (pub. order) within the meta-analysis (z-scaled by meta-analysis), and origin of corresponding author (United States vs. all other countries). Study effects (upper four rows) are estimated without interaction effects. The latter were estimated in a hierarchically well-formulated model, whose main effects (i.e., the same factors that appear above) are omitted. Studies are classified as in Fig. 1.

Inflated effects published by authors in Asian and developing countries had been observed before and are usually explained as the effect of editorial biases: Results from these countries tend to be rejected, unless they report extraordinary findings (36, 37). Inflated effects from the United States, however, cannot be explained by the same mechanism. If anything, English-language journals—whose editors and reviewers are in large proportion American—are sometimes considered biased in favor of North American research (38, 39). Therefore, the US-linked overestimations that we observed among BE/BB studies are very unlikely to be caused by a greater difficulty in publishing negative results. They are more likely the result of questionable methodological choices, such as selectively reporting results of exploratory analyses and "vibration of effects" where different results are obtained depending on what analysis is used (12, 27).

All scientists have to make choices throughout a research project, from formulating the question to submitting results for publication. However, the flexibility with which such choices are made is (by definition) inversely proportional to the level of scholarly consensus within a discipline, which in turn is at least partly determined by the complexity of subject matter and related methodological obstacles (17, 40). Such obstacles are likely to increase, on average, when moving from the physical to the medical and social sciences (i.e., scientific disciplines that are intuitively considered softer). When choices are less rigidly determined by theory, they are more likely to be influenced, consciously or unconsciously, by scientists' own beliefs, expectations, and wishes, and the most basic scientific desire is that of producing an important research finding (e.g., 41, 42). Expectancy effects and similar biases may occur in all fields, but behavioral methodologies have long been known to run a particularly high risk (e.g., 15, 16, 41).

We failed to observe any "US effect" at all among genetic research, thus contradicting observations made by an independent, smaller study in genetic epidemiology (23). This suggests that the prevalence of this and similar patterns needs to be assessed field by field. However, previous observations suggest that a US propensity to report positive and statistically significant results is not limited to particular methodologies and may cut across many disciplines (23–25, 43).

Where would a possible US predisposition toward reporting strong results come from? A complete explanation would probably invoke a combination of cultural, economic, psychological, and historical factors, which at this stage are largely speculative. Our preferred hypothesis is derived from the fact that researchers in the United States have been exposed for a longer time than those in other countries to an unfortunate combination of pressures to publish and winner-takes-all system of rewards (20, 22). This condition is believed to push researchers into either producing many results and then only publishing the most impressive ones, or to make the best of what they got by making them seem as

Table 2. Predictors of over/underestimation, world regions

Predictor	Nonbehavioral ($k = 40$, $n = 566$)	Behavioral, all ($k = 42$, $n = 608$)	Biobehavioral ($k = 20$, $n = 308$)	Behavioral ($k = 22$, $n = 300$)
(Intercept)	0.41 [0.37, 0.44]	0.58 [0.55, 0.60]	0.54 [0.50, 0.58]	0.61 [0.54, 0.63]
Asia	0.01 [−0.01, 0.05]	0.00 [−0.08, 0.04]	0.05 [−0.02, 0.11]	−0.10 [−0.22, −0.01]
EU15	0.01 [−0.01, 0.07]	−0.05 [−0.09, −0.03]	−0.03 [−0.06, 0.02]	−0.07 [−0.11, −0.04]
Other	0.03 [0.00, 0.09]	−0.01 [−0.05, 0.01]	−0.04 [−0.12, −0.02]	0.00 [−0.04, 0.03]
Study size (SE)	0.42 [0.27, 0.52]	0.11 [0.08, 0.24]	0.18 [0.09, 0.29]	0.08 [0.03, 0.33]
Pub. order	0.02 [0.00, 0.03]	0.00 [−0.01, 0.02]	0.01 [0.00, 0.05]	0.00 [−0.01, 0.02]
Asia*SE	0.28 [−0.12, 0.62]	0.05 [−0.28, 0.41]	−0.13 [−0.49, 0.24]	0.64 [−0.18, 1.30]
EU15*SE	0.16 [−0.35, 0.44]	0.37 [0.17, 0.50]	0.16 [−0.16, 0.35]	0.51 [0.27, 0.77]
Other*SE	0.14 [−0.33, 0.47]	0.02 [−0.13, 0.22]	0.27 [0.00, 0.55]	−0.06 [−0.29, 0.25]
Asia*pub. order	0.01 [−0.03, 0.04]	−0.09 [−0.15, −0.03]	−0.07 [−0.14, −0.01]	−0.06 [−0.19, 0.09]
EU15*pub. order	0.03 [0.00, 0.06]	0.02 [0.00, 0.05]	0.04 [0.01, 0.09]	0.02 [−0.03, 0.05]
Other*pub. order	0.02 [−0.04, 0.06]	0.03 [−0.04, 0.04]	0.01 [−0.05, 0.06]	0.06 [0.00, 0.10]

Likelihood of a primary study within a meta-analysis to deviate from the summary effect size (deviation score, i.e., double square root-transformed absolute value of deviation values in Fig. 1), depending on its size (SE), its chronological order of appearance (pub. order) within the meta-analysis (z-scaled by meta-analysis), and the geographical origin of its corresponding author (United States is the reference category). Study effects (upper four rows) are estimated without interaction effects. The latter were estimated in a hierarchically well-formulated model, whose main effects (i.e., the same factors that appear above) are omitted. Studies are classified as in Fig. 1.

Table 3. Predictors of over/underestimation in favor of H1, United States vs. all other countries

Predictor	Behavioral, all ($k = 42$, $n = 608$)	Biobehavioral ($k = 20$, $n = 308$)	Behavioral ($k = 22$, $n = 300$)
(Intercept)	−0.03 [−0.07, 0.04]	0.02 [−0.08, 0.11]	−0.06 [−0.13, 0.04]
United States vs. rest	−0.11 [−0.18, −0.05]	−0.10 [−0.22, 0.02]	−0.11 [−0.20, −0.02]
Expectation (dummy)	−0.06 [−0.11, 0.00]	−0.10 [−0.19, −0.01]	−0.04 [−0.11, 0.07]
(United States vs. rest)*expect.	0.13 [0.04, 0.20]	0.15 [0.01, 0.28]	0.11 [−0.02, 0.20]
SE	0.21 [0.06, 0.29]	0.14 [−0.02, 0.29]	0.24 [0.00, 0.38]
Pub. order	0.01 [−0.02, 0.02]	−0.02 [−0.06, 0.00]	0.02 [0.00, 0.05]

Likelihood of a primary study within a meta-analysis to deviate from the summary effect size (deviation score, i.e., double square root-transformed absolute value of deviation values in Fig. 1), depending on the primary study's size (SE), its chronological order of appearance (pub. order) within the meta-analysis (z-scaled by meta-analysis), geographical origin of corresponding author, and whether the experimental hypothesis tested in the meta-analysis predicted a protective effect (OR <1) or not (an "expectation" dummy variable). A positive interaction term with this variable indicates a greater likelihood to deviate from the summary effect size in the direction predicted by the experimental hypothesis. Studies are classified as in Fig. 1.

important as possible, through post hoc analyses, rehypothesizing, and other more or less questionable practices (e.g., 10, 13, 22, 26). Such a pattern of modulating forces may gradually become more prevalent also in other countries currently and in the near future (18, 20, 21).

Previous observations that positive results are more frequent in the United States, and particularly in the most academically productive states, could be interpreted as evidence of methodological superiority: A greater concentration of talent and research resources might increase the ability to formulate correct hypotheses as well as the statistical power to detect supports for them (17, 25, 26). This study does not favor such interpretation, because it controlled both for choice of hypotheses and for a proxy of quality and statistical power (study size). Even assuming that studies from the United States are methodologically superior in other ways, it would be unclear why this superiority should manifest itself as a greater likelihood to report extreme findings, which favor the experimental hypothesis, and only so in behavioral research.

The patterns observed in this study seem to be different from, and independent of, the classic file-drawer problem, which is often incorrectly equated with small-study effects (inverse correlation between study size and magnitude of reporting) (44). Even assuming that small-study effects in our sample were linked to a file-drawer effect, this would suggest the problem to be lower for the United States and stronger in nonbehavioral studies (Tables 1–4 and Fig. 2).

How these results reflect the situation of other fields and countries remains to be established. Our sampling criteria were primarily dictated by the need to collect reliable data. The meta-analyses included here, therefore, may not be representative of all behavioral and biological (or even just genetic) research, let alone other fields. Moreover, the study was underpowered to detect anything but broad continental differences, and Asian countries were particularly underrepresented among behavioral studies. We suspect that, given sufficient statistical power, these and other countries would reveal their own individual biases, probably limited to specific fields and/or periods in time.

Methods

Meta-Analyses Sampling. The Web of Science database was searched for studies that had the terms "meta-analy*" or "meta analy*" or "systematic review" or "metaanaly*" in the title, and "ratio" or "OR" in the abstract (to try to retrieve papers reporting odds ratio). The search was restricted to document types "article" or "review" and to the Web of Science categories of Psychiatry (PS) and Genetics & Heredity (GH). All meta-analyses reporting primary study-effect sizes in forest plots and having between 10 and 20 primary studies in at least one forest plot were selected for candidate inclusion in the study. This limit was dictated by the need to have balanced numbers of primary studies in each meta-analysis, which was particularly important owing to the hierarchical nature of the data (primary studies, nested in meta-analyses) (33, 34). Only one forest plot was taken from each meta-analysis, usually the first to appear in the study, unless another one had a larger sample size. Obviously, this selection was entirely blind to the country of primary studies' authors.

We initially searched for all meta-analyses published in 2010 and 2011. This gave a sample of 79 apparently usable meta-analyses. Fifteen of these, however, did not report outcomes in odds ratios or in metrics that could be converted to odds ratio (i.e., correlation coefficient, Cohen's d, and Hedges' g). Having verified that that these studies (with outcomes in risk-ratio, mean difference, weighted mean difference, and proportion) could not be combined with the rest, and having noticed the expected effects in the rest of the studies, we expanded the sample to include all usable meta-analyses from 2009 and usable meta-analyses from 2012 up to reaching a total sample of 82 meta-analyses (Table S9 shows meta-analyses reference and key characteristics; primary-study data are available from the corresponding author).

Table 4. Predictors of over/underestimation in favor of H1, world regions

Predictor	Behavioral, all ($k = 42$, $n = 608$)	Biobehavioral ($k = 20$, $n = 308$)	Behavioral ($k = 22$, $n = 300$)
(Intercept)	−0.15 [−0.20, −0.08]	−0.08 [−0.19, 0.03]	−0.17 [−0.24, −0.07]
Asia	0.13 [−0.03, 0.30]	0.18 [0.00, 0.37]	−0.03 [−0.32, 0.29]
EU15	0.09 [0.02, 0.16]	0.09 [−0.05, 0.23]	0.08 [0.00, 0.17]
Other	0.22 [0.07, 0.33]	0.02 [−0.15, 0.19]	0.35 [0.13, 0.50]
Expectation	0.08 [0.00, 0.14]	0.05 [−0.06, 0.16]	0.08 [−0.02, 0.18]
Asia*expectation	−0.20 [−0.38, 0.01]	−0.28 [−0.50, −0.06]	0.01 [−0.34, 0.37]
EU15*expectation	−0.12 [−0.20, −0.03]	−0.16 [−0.32, −0.01]	−0.09 [−0.19, 0.04]
Other*expectation	−0.24 [−0.33, −0.05]	−0.03 [−0.22, 0.16]	−0.35 [−0.50, −0.09]
SE	0.21 [0.07, 0.30]	0.15 [0.00, 0.31]	0.24 [−0.01, 0.38]
Pub. order	0.01 [−0.01, 0.03]	−0.02 [−0.06, 0.00]	0.03 [0.00, 0.06]

Likelihood of a primary study within a meta-analysis to deviate from the summary effect size (deviation score, i.e., double square root-transformed absolute value of deviation values in Fig. 1), depending on the primary study's size (SE), its chronological order of appearance (pub. order) within the meta-analysis (z-scaled by meta-analysis), the geographical origin of its corresponding author (United States is the reference category), and whether the experimental hypothesis tested in the meta-analysis predicted a protective effect (OR<1) or not (an "expectation" dummy variable). A positive interaction term with this variable indicates a greater likelihood to deviate from the summary effect size in the direction predicted by the experimental hypothesis. Studies are classified as in Fig. 1.

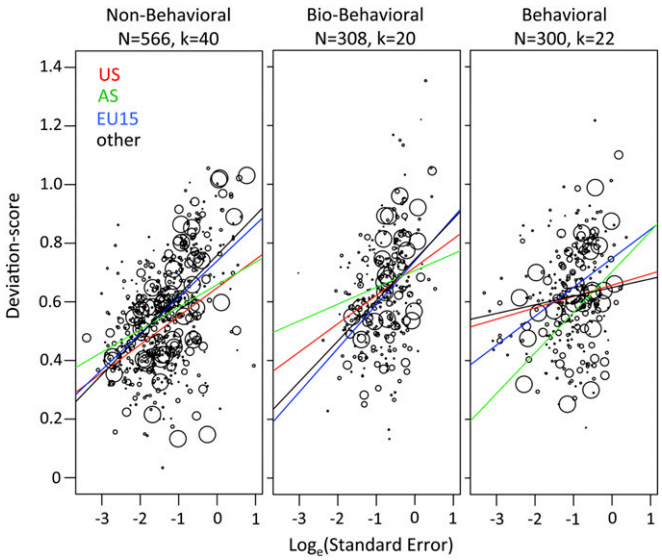

Fig. 2. Deviation score (i.e., double square root-transformed absolute value of data points shown in Fig. 1) plotted against study size [log(SE)], partitioned by type of study. The value of 0 corresponds to a perfect matching between primary study and the summary effect size calculated from the study's meta-analysis, using a random-effects model. Size of circle is inversely proportional to a study's order of appearance (chronological order of publication) in its meta-analysis: larger circles indicate first studies. Nonweighted regression lines were added to illustrate trends for studies of different geographical origin (based on corresponding author). Discipline classification follows the Web of Science system and is based on journal of publication. AS, China, Japan, South Korea, Taiwan, Singapore, and India; EU15, European Union-15 countries; US, United States.

Methodological classification followed previous indications that studies reporting behavioral outcomes might express greater biases than studies combining behavioral and more objective outcomes (i.e., a physical or chemical parameter, or other unambiguous outcome such as death or physical diseases status) (15–17, 41). Meta-analyses were classified with a similar logic, focusing on the particular forest plot that had been coded. This part of the coding was done by D.F., subsequent to the data collection by assistant Leeanne Wood, and was blind to results. In principle, classification could be based exclusively on the outcome, or on the combination of outcome and treatment (i.e., meaning that studies assessing a physiological treatment on behavioral outcome would be classified as BE in the former case and BB in the latter). In practice, this created ambiguities only in six meta-analyses, all from PS, so coding followed the latter logic, deemed more intuitive. All meta-analyses in GH were classifiable as NB, whereas in PS one meta-analysis was NB, 22 (51%) were purely BE, and 20 were BB.

The Web of Science classification is based on journal and includes 151 nonexclusive categories, so the two categories used for sampling overlapped with each other, as well as with 26 other categories. In our final sample, the NB meta-analyses were classified in the following Web of Science categories: genetics & heredity (n = 39), biochemistry molecular biology (n = 8), toxicology (n = 8), immunology (n = 6), cell biology (n = 2), oncology (n = 2), and nine other categories (n = 1 each). The BB meta-analyses were classified in psychiatry (n = 20), clinical neurology (n = 5), neurosciences (n = 4), psychology (n = 4), psychology multidisciplinary (n = 3), surgery (n = 3), pharmacology pharmacy (n = 2), and seven other categories (n = 1 each). The BE meta-analyses were classified in psychiatry (n = 22), psychology clinical (n = 9), psychology (n = 8), geriatrics gerontology (n = 2), gerontology (n = 2), and two other categories (n = 1 each).

Data Collection. From each included forest plot, research assistant Leeanne Wood, who was blind to the hypothesis, recorded effect sizes and measures of precision (usually confidence interval), study ID, and the overall pooled effect size. Bibliographic data available in the Web of Science for each primary study was then retrieved. If this was unavailable, the parameters that were key to the study (i.e., year of publication, number of authors, and corresponding address) were obtained by Web search. In the few cases when none of this information could be obtained (n = 9 for authorship and n = 13 for year) missing values were replaced by approximations of the average values of the rest of the sample (i.e., six authors, year 2003).

Data Elaboration. Primary studies' outcomes reported as odds ratios were included directly. Those reported as Cohen's d, Hedges' g, and correlation coefficient were converted to odds ratio following standard transformations (45). Values for confidence intervals were also transformed, and then SEs were obtained by subtracting the upper confidence interval from the mean value and dividing by 1.96. The same calculation was performed using the lower confidence interval, to double-check the accuracy of the data. To assess how much each primary study had under- or overestimated the effect size, and to verify whether such deviations would have supported researchers' own expectations, we used the following methods.

Deviation score. This measure is based on that used in a previous metameta-analysis (23), which we modified to encompass general cases. Using the reported primary study outcomes, we recalculated the summary effect size for each meta-analysis using the metagen function in the metafor R package (46). Then each primary effect size was divided by the summary effect size of its respective meta-analysis, producing a value varying between 0 and infinity, centered on 1.

The ratio values were then log-transformed (base 10), to produce a set of values ranging from –inf to +inf, and centered on 0. Negative values indicate underestimation and positive values overestimation. These are the values shown in Fig. 1. To obtain the deviation score used in the analyses, we "folded" the distribution by taking the absolute value and square-root-transformed these values twice to achieve normality (checked by examining the histogram and values of skewness and kurtosis). In sum, the deviation score of effect size d from primary study i in meta-analysis j was calculated as

$$\sqrt[4]{\left|\mathrm{Log}_{10}\left(\frac{d_{ij}}{\overline{d}_j}\right)\right|},$$

where $\overline{d}_j$ is the summary effect size of meta-analysis j.

Expectation factor. Meta-analyses were classified according to whether the tested effect was expected to be protective (i.e., odds ratio <1) or not (odds ratio >1). Dummy variables were created by examining the forest plot or reading the abstract and eventually the full text to identify the direction of expected effects (i.e., of the hypothesis 1). This dummy variable should, in theory, not be a significant predictor of unfolded deviation scores (i.e., those in Fig. 1 and Fig. S1). A significant positive effect of this variable suggests a propensity to over- or underestimate effects in the direction that rejects the null (i.e., that favors treatment against control groups, or shows a significant association between two variables, in the direction hypothesized). Interaction effects between expectation factor and country, therefore, test for the propensity of a country to report a stronger effect in support of the hypothesis relative to the reference country (i.e., United States vs. all other countries, or different geographic areas vs. United States, depending on the test).

In most cases, meta-analyses in PS contrasted a treatment and control group and indicated clearly what the tested hypothesis is, often in the forest plot itself. This allowed the expectation factor to be coded objectively. Most GH meta-analyses, however, assessed correlations between gene variants and medical conditions. Inferring the "preferred" direction in such cases is a subjective and unreliable exercise, so we only tested the expectation factor among BE and BB studies.

Robustness Analyses. The summary estimates for each meta-analysis were recalculated twice, assuming fixed and random effects, respectively. In the main text we report results obtained with random-effects values, and in *Supporting Information* we show results obtained with the fixed effects as well as the summary estimate that was reported in the original forest plot.

We also tried measuring primary studies' deviations using alternative approaches. In particular, we tried using z-score–like formulas, using meta-analytical summary estimates or simple mean. We called these, respectively, Meta-analysis-scaled deviation score and Z-scaled deviation score and calculated them using the following formulas.

Meta-analysis–scaled deviation score. For effect size d of primary study i in meta-analysis j, it was calculated as

$$\sqrt[4]{\left|\frac{d_{ij}-\overline{d}_j}{s_j}\right|},$$

where $\overline{d}_j$ is the summary effect size of the meta-analysis and s_j is the SD of effect sizes. Square-root transformation was needed to achieve normality (checked by examining the histogram and values of skewness and kurtosis).

We recalculated this value using the fixed-effects and original summary estimate, but for brevity we only present results obtained with random effects.

Z-scaled deviation score. For effect size *d* of primary study *i* in meta-analysis *j*, it was calculated by

$$\frac{d_{ij}-\mu_j}{s_j},$$

where μ_j is the arithmetical mean and s_j is the SD of effect sizes.

Scaled weighting. Following the suggestion of one reviewer, we also examined how centering weights by meta-analysis affected our results. In all our main analyses, studies are weighted by the inverse square of their original SE values based on the assumption that information on the relative sizes of studies across meta-analysis should be maintained. For example, if two studies in two different meta-analyses overestimated the effect to the same extent, but one is 10 times larger than the other, our analyses give more weight to the larger study. To assess how results depended on this assumption, we rescaled primary studies' SEs to the mean SE values for the whole sample. For SE *t* of primary study *i* in meta-analysis *j*, we calculated

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

where $\bar{t}_j$ is the average SE in meta-analysis *j* and $\hat{t}$ is the average of all 1,174 SEs in the sample.

Statistical Analyses. Owing to the Proteus phenomenon (extreme contradictory effects published easier and sooner) and other bias-related patterns, magnitudes of outcomes appearing within a meta-analysis are not necessarily independent of one another (4, 14). It is likely that studies in some meta-analyses might diverge more than in others. To control for this between–meta-analyses effect, we used a two-level generalized linear model, which controlled for random differences in the average level of deviation within meta-analyses (i.e., random intercept at level-2) (33, 34, 35). Unless differently specified, the model followed standard meta-analytical procedures and weighted studies by the inverse of the square of their SE (which in meta-analysis is called inverse-variance weighted regression, or metaregression). Predictors' confidence intervals were assessed by Markov chain Monte Carlo sampling, with 50,000 iterations.

Nonindependence of data points was further avoided by including only one data point from each primary study: If more than one outcome from the same study appeared in the forest plot, only one of them was selected for the final analyses, using a random number generator.

In *Results* we present results progressively: We first overview the data as it appears in figures, then report, in the text, values obtained adding one covariate at a time to the models, and finally report the multiple-regression values in tables, separating main effects and interactions. All effect sizes are given with their 95% confidence interval, unless differently specified.

ACKNOWLEDGMENTS. This work was supported by Leverhulme Early-Career Fellowship ECF/2010/0131 (to D.F.). J.P.A.I. was supported by an unrestricted gift from Sue and Bob O'Donnell to the Stanford Prevention Research Center.

1. Song F, et al. (2010) Dissemination and publication of research findings: An updated review of related biases. *Health Technol Assess* 14(8):iii, ix–xi, 1–193.
2. Rosenthal R (1979) The file drawer problem and tolerance for null results. *Psychol Bull* 86(3):638–641.
3. Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. *Nat Genet* 29(3):306–309.
4. Ioannidis JPA, Trikalinos TA (2005) Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *J Clin Epidemiol* 58(6):543–549.
5. Schooler J (2011) Unpublished results hide the decline effect. *Nature* 470(7335): 437–437.
6. Ioannidis JPA (2008) Why most discovered true associations are inflated. *Epidemiology* 19(5):640–648.
7. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG (2004) Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA* 291(20):2457–2465.
8. Chan AW, Krleza-Jerić K, Schmid I, Altman DG (2004) Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 171(7): 735–740.
9. Chan AW, Altman DG (2005) Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *BMJ* 330(7494):753–756.
10. Fanelli D (2009) How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE* 4(5):e5738.
11. Stroebe W, Postmes T, Spears R (2012) Scientific misconduct and the myth of self-correction in science. *Perspect Psychol Sci* 7(6):670–688.
12. John LK, Loewenstein G, Prelec D (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol Sci* 23(5):524–532.
13. Fang FC, Steen RG, Casadevall A (2012) Misconduct accounts for the majority of retracted scientific publications. *Proc Natl Acad Sci USA* 109(42):17028–17033.
14. Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2(8): e124.
15. Rosenthal R (1976) *Experimenter Effects in Behavioural Research. Enlarged Edition* (Irvington Publishers, New York).
16. Burghardt GM, et al. (2012) Perspectives - Minimizing observer bias in behavioral studies: A review and recommendations. *Ethology* 118(6):511–517.
17. Fanelli D (2010) "Positive" results increase down the hierarchy of the sciences. *PLoS ONE* 5(4):e10068.
18. Qiu J (2010) Publish or perish in China. *Nature* 463(7278):142–143.
19. Osuna C, Cruz-Castro L, Sanz-Menéndez Luis (2011) Overturning some assumptions about the effects of evaluation systems on publication performance. *Scientometrics* 86:575–592.
20. De Rond M, Miller AN (2005) Publish or perish: Bane or boon of academic life? *J Manage Inq* 14(4):321–329.
21. de Meis L, Velloso A, Lannes D, Carmo MS, de Meis C (2003) The growing competition in Brazilian science: Rites of passage, stress and burnout. *Braz J Med Biol Res* 36(9): 1135–1141.
22. van Dalen HP, Henkens K (2012) Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. *J Am Soc Inf Sci Technol* 63(7):1282–1293.
23. Munafò MR, Attwood AS, Flint J (2008) Bias in genetic association studies: Effects of research location and resources. *Psychol Med* 38(8):1213–1214.
24. Doucouliagos H, Laroche P, Stanley TD (2005) Publication bias in union-productivity research? *Relations Industrielles-Industrial Relations* 60(2):320–347.
25. Fanelli D (2012) Negative results are disappearing from most disciplines and countries. *Scientometrics* 90(3):891–904.
26. Fanelli D (2010) Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PLoS ONE* 5(4):e10271.
27. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22(11):1359–1366.
28. Francis G (2012) Too good to be true: publication bias in two prominent studies from experimental psychology. *Psychon Bull Rev* 19(2):151–156.
29. Ferguson CJ, Brannick MT (2012) Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychol Methods* 17(1):120–128.
30. Fiedler K (2011) Voodoo correlations are everywhere—not only in neuroscience. *Perspect Psychol Sci* 6(2):163–171.
31. Bakker M, Wicherts JM (2011) The (mis)reporting of statistical results in psychology journals. *Behav Res Methods* 43(3):666–678.
32. Fava GA (2007) Financial conflicts of interest in psychiatry. *World Psychiatry* 6(1): 19–24.
33. Raudenbush SW, Spybrook J, Liu X, Congdon R (2005) *Optimal Design for Longitudinal and Multilevel Research* (Univ of Chicago Press, Chicago), Version 1.55.
34. Scherbaum CA, Ferreter JM (2009) Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organ Res Methods* 12(2): 347–367.
35. Nezlek JB (2011) *Multilevel Modeling for Social and Personality Psychology* (Sage, London).
36. Yousefi-Nooraie R, Shakiba B, Mortaz-Hejri S (2006) Country development and manuscript selection bias: A review of published studies. *BMC Med Res Methodol* 6:37.
37. Matías-Guiu J, García-Ramos R (2011) Editorial bias in scientific publications. *Neurologia* 26(1):1–5.
38. Link AM (1998) US and non-US submissions: An analysis of reviewer bias. *JAMA* 280(3):246–247.
39. Lynch JR, et al. (2007) Commercially funded and United States-based research is more likely to be published; good-quality studies with negative outcomes are not. *J Bone Joint Surg Am* 89(5):1010–1018.
40. Fanelli D, Glänzel W (2013) Bibliometric evidence for a hierarchy of the sciences. *PLoS ONE* 8(6):e66938.
41. van Wilgenburg E, Elgar MA (2013) Confirmation bias in studies of nestmate recognition: a cautionary note for research into the behaviour of animals. *PLoS ONE* 8(1): e53548.
42. Jeng M (2006) A selected history of expectation bias in physics. *Am J Phys* 74(7): 578–583.
43. Fanelli D (2012) When East meets West... does bias increase? A preliminary study on South Korea, United States and other countries. *Proceedings of the 8th International Conference on Webometrics, Informetrics and Scientometrics and 13th COLLNET Meeting*, eds Choi H-N, et al. (Korea Institute of Science and Technology Information, Seoul), pp 218–223.
44. Ioannidis JPA (2008) Interpretation of tests of heterogeneity and bias in meta-analysis. *J Eval Clin Pract* 14(5):951–957.
45. Lipsey MW, Wilson DB (2001) *Practical Meta-Analysis* (Sage Publications, Thousand Oaks, CA).
46. Viechtbauer W (2010) Conducting meta-analyses in {R} with the {metafor} package. *J Stat Softw* 36(3):1–48.