Uncovering Null Effects in Null Fields: The Case of Homeopathy

Edgar Erdfelder[1], Juliane Nagel[2], Daniel W. Heck[3], and Nils Petras[1]

[1]Department of Psychology, University of Mannheim, Mannheim, Germany

[2]Department of Clinical Psychology, Department of Addiction Behavior and Addiction Medicine, Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany

[3]Department of Psychology, University of Marburg, Marburg, Germany

Author Note

Edgar Erdfelder https://orcid.org/0000-0003-1032-3981

Juliane Nagel https://orcid.org/0000-0002-5310-8088

Daniel W. Heck https://orcid.org/0000-0002-6302-9252

Nils Petras https://orcid.org/0000-0001-9528-2298

All data, analysis scripts, and computer code in R required for reproducing the results reported in this work are provided online on the Open Science Framework (OSF; https://osf.io/wuq2h/). We have no conflicts of interest to disclose. This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), grant GRK 2277, Research Training Group *Statistical Modeling in Psychology* (SMiP). The current research was initiated by a discussion in the Mannheim ReproducibiliTea meeting which is part of the ReproducibiliTea network (https://reproducibilitea.org/). Correspondence can be addressed to any of the authors EE (edgar.erdfelder@uni-mannheim.de), JN (juliane.nagel@zi-mannheim.de), DWH (daniel.heck@uni-marburg.de), or NP (nils.petras@uni-mannheim.de). Postal address: Edgar Erdfelder, Cognition and Individual Differences Lab, Department of Psychology, School of Social Sciences, University of Mannheim, A5, Room B 118, 68159 Mannheim, Germany

Abstract

Objective: Sigurdson, Sainani, and Ioannidis (this journal) discussed homeopathy as a prototypical example of a "null field" where true effects are nonexistent and positive effect sizes reflect bias only. Based on a sample of published randomized placebo-controlled trials, they observed a surprisingly large effect in favor of homeopathy (Hedges' $g = 0.36$). In this comment, we propose selective publication of significant results as a parsimonious explanation of the overall bias evident in this field.

Study Design: We re-analyzed the data of Sigurdson and collaborators using a meta-analytic mixture model that accounts for selective publishing with two parameters only, (1) the true homeopathy effect and (2) the proportion of results published only when statistically significant in the predicted direction.

Results: The mixture model fitted the data. As expected, the estimate of the true homeopathy effect reduces to almost zero ($\hat{d} = 0.05$, 95% CI: [-0.05 - 0.16]) when taking selective publishing into account.

Conclusion: Inclusion of effect size measures adjusting for selective publication practices should become routine practice in meta-analyses. Null fields not only provide useful benchmarks for the overall bias evident in a field. They are also important for testing explanations of this bias and validating adjusted effect size measures.

Keywords: meta-analysis, effect size measures, selective publishing, null fields, homeopathy

Introduction

In a recent contribution to this journal, Sigurdson et al. (2023) proposed "null fields" as negative controls for scientific results. "Null fields" have been characterized as fields where no true effect can be discovered according to accepted standards of scientific reasoning (Ioannidis, 2005). Hence, if effects emerge in a null field, they reflect nothing but the composite influence of systematic biases.

In their meta-analytic study, Sigurdson and collaborators investigated homeopathy as a candidate for such a null field. The mechanisms by which homeopathy supposedly works have been argued to be physically impossible and thus inconsistent with basic scientific principles (Grimes, 2005). Hence, the true effect of homeopathy compared to placebo controls is expected to be zero.

Using 50 published randomized controlled trials (RCTs) sampled from the most frequently cited meta-analyses of homeopathic treatments, the authors observed a remarkable effect in favor of homeopathy, more precisely, Hedges' $g = 0.36$ (95% CI: [0.21 - 0.51]). For future research, they recommended "calibrated comparisons to the expected level of bias rather than to the number that represents a null effect" (Sigurdson et al., 2023). Hence, in the field of homeopathy, $g = 0.36$ reflects the typical bias that may be taken as "the average transformation of the null".

We agree that $g = 0.36$ is a substantial effect that provides a useful benchmark of overall bias in the field of homeopathy. In this comment, we aim at a parsimonious explanation of this bias. In what follows, we first discuss selective publication practices as potential sources of bias. We then make use of an alternative effect-size measure, termed the mixture-model measure, that takes selective publishing into account. To foreshadow, this refined measure results in an estimate of virtually zero for the true homeopathy effect while the underlying statistical model describes the meta-analytic data well. Finally, we discuss implications for future research.

The problem of selective publication conditional on significance

An ideal meta-analysis of the average homeopathy effect would include (a) a representative sample of all homeopathy RCTs ever conducted and (b) only studies

analyzed properly, using a single statistical assessment of a pre-defined dependent variable irrespective of the outcome. Unfortunately, published research rarely follows this ideal (e.g., Greve et al., 2013; Simmons et al., 2011; Stanley et al., 2021). Most likely, this also holds for Sigurdson et al.'s prototypical homeopathy studies. While some included results may conform to the ideal and were published irrespective of the outcome, others were published only when statistically significant at $\alpha = .05$ while non-significant effects remained unpublished. This gives rise to a positive bias in the results that survive the selection process. Hence, published effects are not anymore representative of the true underlying effect.

The possible reasons are manifold. Selective publishing is not a single bias but a collection of biases that operate at different levels. Some biases originate in researchers, for example, when they refrain from submitting their research for publication once they observe a non-significant outcome. Others may *p*-hack their data until they discover a significant outcome (Simmons et al., 2011; Simonsohn et al., 2014), for example, by trying different dependent variables ("multiple testing"), analyzing different subsets of the data ("data peeking"), modifying data exclusion criteria ("data trimming") or changing the statistical analysis method ("model switching") when the initial analysis did not produce the desired result (Simmons et al., 2011). Still other selective publication practices originate in publishers, editors or reviewers who may be reluctant to publish non-significant outcomes. Although considerable effort has been directed towards identifying the specific reasons for selective publication of significant results (e.g., Simonsohn et al., 2014), several authors have argued that inferring these reasons from published results alone is difficult if not impossible (e.g., Erdfelder & Heck, 2019; Ulrich & Miller, 2015).

Fortunately, correcting meta-analytic effect size measures for distortions due to selective publishing does not require knowledge of all sources responsible for this bias. It suffices to set up a statistical model that describes the composition of published effect sizes correctly. For mean comparisons of continuous outcome variables between two randomized groups, non-adjusted Hedges' g effect size estimates directly translate in the

well-known Student's t-statistic,

$$t_{(df)} = g \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}, \tag{1}$$

where n_1 and n_2 denote the sample sizes of the two groups and $df = n_1 + n_2 - 2$ the degrees of freedom of the respective t-statistic. Making use of this fact, Ulrich et al. (2018, p. 68-71) proposed a simple finite mixture model of $t_{(df)}$ under selective publishing. According to this model, the distribution of published effect sizes in a field (and the corresponding $t_{(df)}$-distribution) consists of two components: An unknown proportion p_u of results published unconditionally, that is, irrespective of the outcome of the t-test, and a complementary proportion $p_c = 1 - p_u$ published conditional on the t-test being significant (e.g., at $\alpha = .05$). Hence, the resulting distribution must be a two-component mixture of a standard and a truncated (noncentral) t-distribution. The model has two free parameters: the true effect d in the field and the probability of selective publishing, p_c. For formal details of the mixture model, parameter estimation, and goodness-of-fit testing, we refer to supplemental materials available at the Open Science Framework (OSF, see https://osf.io/wuq2h/).

In the following, we apply this model to the homeopathy data of Sigurdson et al. (2023). Of course, alternative meta-analytic tools to cope with selective publishing exist (for reviews, see Page et al., 2023; Ulrich et al., 2018). Yet, the mixture model has several advantages, especially when applied to null fields. First, the mixture model provides maximum likelihood (ML) estimates $\hat{d}$ of the true effect d decontaminated from distortions due to selective publishing. Second, it does so without requiring assumptions why this bias occurs. It simply accounts for an unknown proportion p_c of results published only when significant, being agnostic about underlying reasons. Third, the mixture model is a fixed-effects model; it assumes that the true effect size is an unknown but fixed constant d. This is perfectly in line with the prediction for null fields. Null fields, by definition, are characterized by a fixed true effect of $d = 0$. Fourth, using the mixture model, the null hypothesis $H_0 : d = 0$ can be tested statistically using a G^2 likelihood ratio test with controlled error probabilities. The same applies to tests of $H_0 : p_c = 0$, (i.e, no selective publishing). Based on the null-field hypothesis, we

expect that $H_0 : d = 0$ best fits the homeopathy data. In addition, we would expect a rejection of $H_0 : p_c = 0$ to explain the overall bias $g = 0.36$ as a consequence of selective publishing. Fifth, as detailed in the supplemental materials, a goodness-of-fit test of the mixture model based on the Kolmogorov-Smirnov statistic is easily derived. Using this test, we can evaluate whether the mixture model fits the distribution of observed g effect sizes (or the associated t-statistics, cf. Equation 1). Sixth, the mixture model takes the entire distribution of observed effect sizes into account (not just the statistically significant part) so that a maximum of information available in the data is used for statistical inference.

In sum, the mixture model nicely fits our research goals: It provides the statistical methods required for testing a parsimonious explanation of the bias evident in homeopathy research.

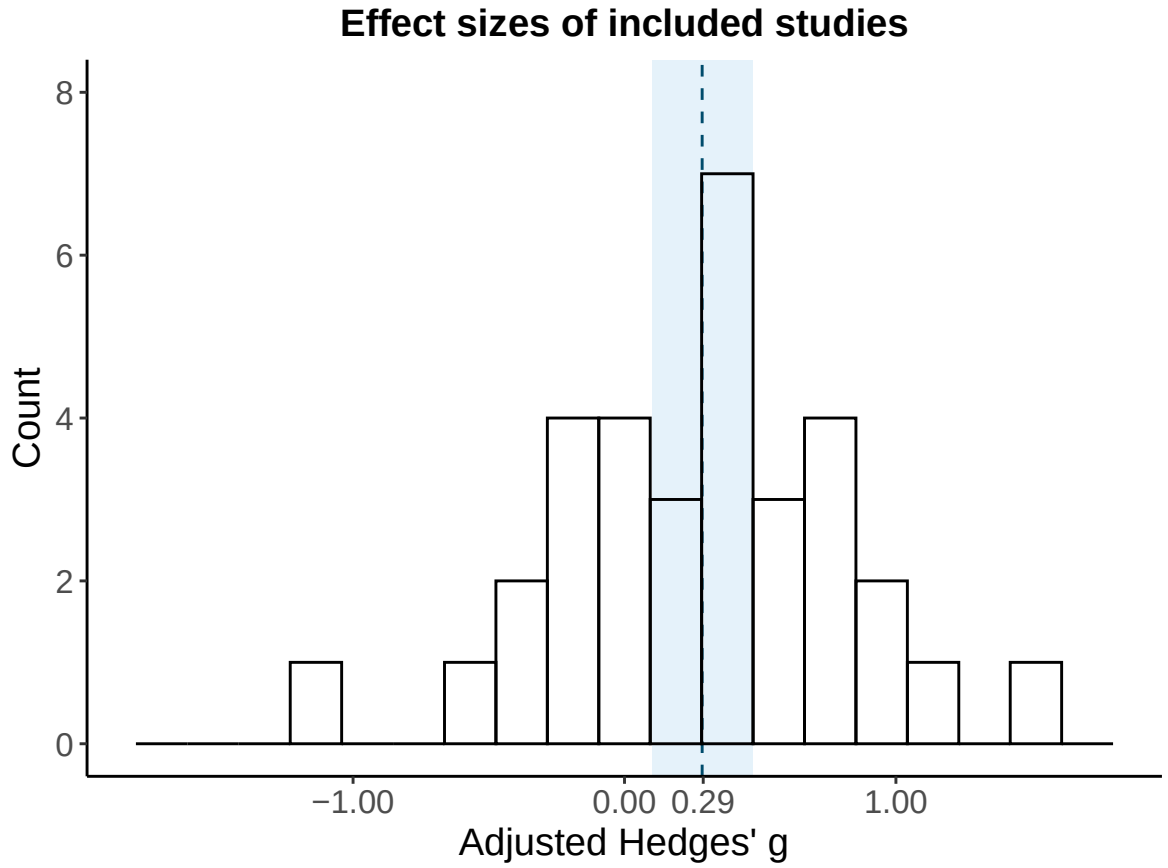
Method

We extracted effect sizes of all 50 RCTs reported by Sigurdson et al. (2023). Since the mixture model applies to t-tests of continuous dependent variables for two independent groups, we limited the relevant effect sizes to standardized mean deviations (SMD) measured in terms of adjusted Hedges' g, thus excluding other effect sizes such as risk ratios (RR) or odds ratios (OR).[1] A total of 37 studies met this criterion. Another 4 studies had to be excluded because outcome measures were obtained from a single group of participants only, leaving us with an effective $k = 33$ of studies that entered in our re-analysis of Sigurdson et al.'s data. The reduced data set of adjusted g estimates along with our analysis scripts employing metamix (Petras & Heck, 2023) is available at the OSF (see https://osf.io/wuq2h/).

[1] While transformations between different effect-size measures are possible in principle (e.g., Chinn, 2000), transformations of RR or OR to t-statistics are problematic because a statistical rationale justifying t-distributions of the transformed values is lacking. Since the t-distribution assumption is crucial for applications of the mixture model, we omitted RR and OR data from our analysis.

Figure 1

Distribution of adjusted Hedges' g effect sizes in the k = 33 studies taken from Sigurdson et al.'s homeopathy data set.



Note. Adjusted Hedges' g is the homeopathy-placebo mean difference divided by the pooled baseline standard deviation weighted by sample sizes and multiplied by the correction factor $J = 1 - 3/(4 \cdot (n_1 + n_2) - 9)$ (see Sigurdson et al., 2023). The dashed vertical line illustrates the mean $g = 0.29$ and the blue shaded area the 95% confidence interval.

Results

Figure 1 illustrates the distribution of adjusted (i.e., J-corrected) g effect sizes for the $k = 33$ relevant studies. Importantly, with a significant mean effect size of $g =$ 0.29 (95% CI: [-0.10 - 0.47]), these studies resemble the overall bias reported for all 50 studies.

Using metamix, we fitted the mixture model to the t-statistics of the $k = 33$ homeopathy RCTs (assuming selective publishing relative to $\alpha = .05$, one-tailed). This resulted in an ML-estimate of $\hat{d} = 0.05$ for the true effect size in the field (SE = 0.05,

95% CI: [-0.05 - 0.16]) which is not significantly different from zero ($G^2(1) = 1.6, p = .30$). In contrast, the ML-estimate for the probability of selective publishing was almost at ceiling, $\hat{p}_c = .86$ (SE = .08, 95% CI: [.63 - .96]), and clearly significant ($G^2(1) = 11.59, p_b < .001$).[2]

Figure 2 illustrates empirical, fitted, and theoretical distributions of t-statistics for one-tailed t-tests under the mixture model assuming $\alpha = .05$. A Kolmogoroff-Smirnov two-sample test reveals that the fitted (blue-shaded) mixture distribution is not significantly different from the observed (black-framed) frequency histogram ($D = .17, p = .27$). Thus, given the current data, there is no reason to reject the mixture model. The figure also illustrates the overall distribution of t-statistics for the entire field of homeopathy RCTs, that is, the union of the blue- and red-shaded areas that represent published and unpublished results, respectively. This distribution resembles a central t-distribution with $d = 0$.

Discussion

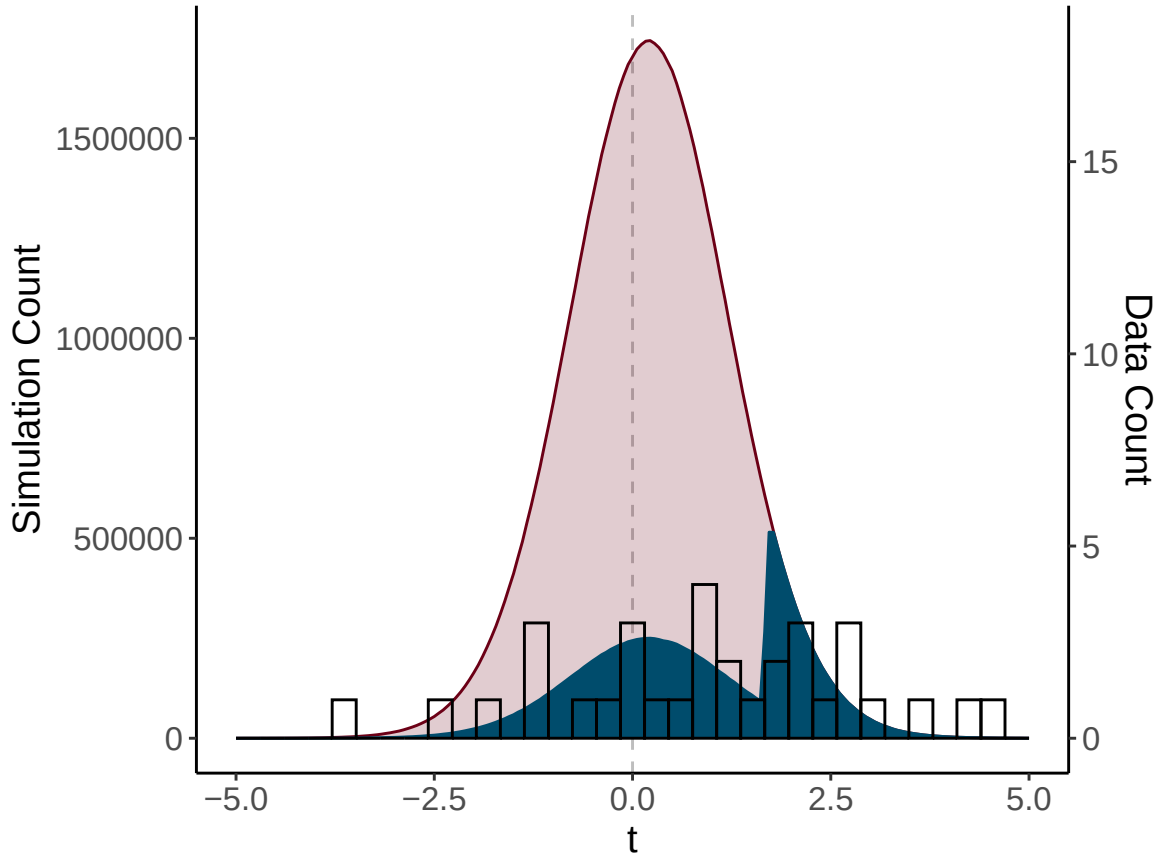
Sigurdson et al. (2023) discussed homeopathy as a null field where true effects are nonexistent. Yet, in a prototypical sample of 50 published placebo-controlled trials, they observed a substantial mean effect of $g = 0.36$ in favor of homeopathy.

We agree that $g = 0.36$ reflects systematic bias and provides a useful benchmark for the overall bias evident in homeopathy research. Our aim was to provide and test a parsimonious explanation of this bias, namely, selective publication of significant results. In particular, the meta-analytic mixture model of Ulrich et al. (2018) is shown to provide a promising framework for the evaluation of alleged null fields. Based on this model, a re-analysis of the homeopathy data with continuous outcome measures yields ML-estimates close to zero for the true treatment effect ($\hat{d} = 0.05$, 95% CI: [-0.05 - 0.16]) and of $\hat{p}_c = .86$ (95% CI: [.63 - .96]) for the probability of selective publishing of significant results. Given that $k = 33$, the power of the G^2 test of $H_0 : d = 0$ exceeds

[2] p_b refers to the chi-bar-square distribution. Because p_c is at the boundary of the parameter space under $H_0 : p_c = 0$, $G^2(1)$ does not follow a standard $\chi^2(1)$-distribution in this case (cf. Molenberghs & Verbeke, 2007).

Figure 2

Observed, fitted, and underlying theoretical distributions of the two-groups t-statistics corresponding to the $k = 33$ studies taken from Sigurdson et al.'s meta-analytical homeopathy data set.



Note. The histogram illustrates the frequency distribution of t-statistics derived from nonadjusted g statistics via Equation (1). The blue-shaded curve illustrates the fitted mixture distribution for one-tailed t-tests assuming selective publishing relative to $\alpha = .05$. The red-shaded area illustrates t-values predicted by the model that did not reach the significance threshold under selective publishing and thus remained unpublished. The blue- and the red-shaded areas combined display the full distribution of t-statistics for the field, as inferred from the mixture model. The dashed vertical line illustrates the mean of zero expected under $H_0 : d = 0$.

54% if $d \geq 0.20$ and 81% if $d \geq 0.40$ under H_1 (see Ulrich et al., 2018, Appendix B, Table A1). Hence, the non-significant outcome of this test supports the claim that homeopathy essentially is a null field. Of similar importance, the model explains the overall bias $g = 0.36$ as a consequence of the large proportion of results published only when significant. The large p_c-estimate also suggests that some form of p-hacking is

likely involved in the field of homeopathy, that is, the number of statistical analyses explored (and typically ignored when the test was insignificant) was supposedly much larger than the number of RCTs conducted.

Importantly, by arguing that selective publication practices can explain the overall bias observed by Sigurdson and collaborators, we do not imply that other biases are not involved in the field of homeopathy. It is indeed conceivable that additional biases operate next to selective publishing (e.g., insufficient blinding procedures, experimenter bias, or additional biases not captured by truncated reporting of $p < .05$ results). Yet, given the mixture model's fit and the insignificant deviation from $H_0 : d = 0$ (see Figure 2), such an assumption appears not necessary for the current data. This of course does not imply that a replication based on a larger meta-analytic data set or applications to alternative null fields will arrive at the same conclusion. Hence, our results should not be taken as the "final word" on the origin of biases in null fields, but rather as a starting point for research that aims at explaining overall biases evident in null fields. This research program should also include additional correction methods with goals similar to those of the mixture model (e.g., Bartoš et al., 2023; Stanley & Doucouliagos, 2022; Stanley et al., 2017; Stanley et al., 2021).[3]

Given the current evidence, we conclude that inclusion of meta-analytic measures adjusting for selective publishing should be routine practice in meta-analyses of published research. Research domains likely to be null fields are very valuable not only for generating benchmarks of overall bias, but also for validating bias-corrected measures. Because our re-analysis of Sigurdson et al.'s homeopathy data uncovered the expected true effect $d = 0$ for a null field, it can be seen as an example that successfully validates the mixture model for the homeopathy field, assuming that this bias is largely due to selective publishing. Domains in which the true effect is supposedly $d > 0$ will of

[3] For comparison purposes, we analyzed the $k = 33$ homeopathy studies also using Robust Bayesian Meta Analysis (RoBMA, cf. Bartoš et al., 2023). Despite considerable discrepancy in the assumptions underlying RoBMA and the mixture model, the adjusted effect estimates are very similar (RoBMA: $\hat{\mu} = 0.07$, 95%CI: -0.04 – 0.36).

course also benefit from such approaches because conventional meta-analytic measures generally tend to be biased to some degree, at least if p_c exceeds zero.

Depending on research questions and conventions, the correction method most appropriate for a field may differ between domains. We therefore favor methods that provide goodness-of-fit tests for the proposed statistical models. Compared to methods that lack model tests, users of testable models have a more solid empirical basis for deciding whether a correction method is appropriate for a domain or not. User friendly software – such as the metamix R package (Petras & Heck, 2023) that accompanies the present publication – will facilitate use of such meta-analytic tools in the future.

Finally, we point out that the mixture model, like the method of overall bias assessment (cf., Sigurdson et al., 2023), can in principle be generalized beyond t-tests to other types of tests, any number of groups, and any number of dependent variables. The basic idea of two-component mixtures of standard and truncated distributions conditional on significance is quite universal and can be adapted to virtually any test statistic in meta-analytic research. Using such generalized models, the null-field hypothesis can be evaluated irrespective of designs and tests that dominate a research field, provided that selective publishing is a major source of bias in a field.

References

Bartoš, F., Maier, M., Wagenmakers, E. J., Doucouliagos, H., & Stanley, T. D. (2023). Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*, *14*, 99–116. https://doi.org/10.1002/jrsm.1594

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*, 3127–3131.

Erdfelder, E., & Heck, D. W. (2019). Detecting evidential value and p-hacking with the p-curve tool: A word of caution. *Zeitschrift für Psychologie*, *222*, 249–260. https://doi.org/10.1027/2151âĂŞ2604/a000383

Greve, W., Bröder, A., & Erdfelder, E. (2013). Result-blind peer reviews and editorial decisions: A missing pillar of scientific culture. *European Psychologist*, *18*, 286–294. https://doi.org/10.1027/1016âĂŞ9040/a000144

Grimes, D. R. (2005). Proposed mechanisms for homeopathy are physically impossible. *Focus on Alternative and Complementary Therapies*, *17*, 149–155.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

Molenberghs, G., & Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician*, *61*, 22–27. https://doi.org/10.1198/000313007X171322

Page, M. J., Higgins, J. P. T., & Sterne, J. A. C. (2023). Assessing risk of bias due to missing results in a synthesis. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions (Version 6.4)*. Cochran.

Petras, N., & Heck, D. W. (2023). metamix: Meta-analytic effect estimates adjusted for selective publishing conditional on significance [R package]. https://github.com/NilsPetras/metamix

Sigurdson, M. K., Sainani, K. L., & Ioannidis, J. P. (2023). Homeopathy can offer empirical insights on treatment effects in a null field. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2023.01.10

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P -curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. https://doi.org/10.1037/a0033242

Stanley, T. D., & Doucouliagos, H. (2022). Harnessing the power of excess statistical significance: Weighted and iterative least squares. *Psychological Methods*. https://doi.org/10.1037/met0000502

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. A. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*, *36*, 1580–1598. https://doi.org/10.1002/sim.7228

Stanley, T. D., Doucouliagos, H., Ioannidis, J. P. A., & Carter, E. C. (2021). Detecting publication selection bias through excess statistical significance. *Research Synthesis Methods*, *12*, 776–795. https://doi.org/10.1002/jrsm.1512

Ulrich, R., & Miller, J. (2015). P-hacking by post hoc selection with multiple opportunities: Detectability by skewness test? Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General*, *144*, 1137–1145. https://doi.org/10.1037/xge0000086

Ulrich, R., Miller, J., & Erdfelder, E. (2018). Effect size estimation from *t*-statistics in the presence of publication bias: A brief review of existing approaches with some extensions. *Zeitschrift für Psychologie*, *226*(1), 56–80. https://doi.org/10.1027/2151-2604/a000319