Journal of Quantitative Criminology (2020) 36:329–346
https://doi.org/10.1007/s10940-018-9387-8

# Interpreting t-Statistics Under Publication Bias: Rough Rules of Thumb

**Christopher Winship**[1] · **Xiaolin Zhuo**[1]

Published online: 10 July 2018


## Abstract

**Introduction** A key issue is how to interpret t-statistics when publication bias is present. In this paper we propose a set of rough rules of thumb to assist readers to interpret t-values in published results under publication bias. Unlike most previous methods that utilize collections of studies, our approach evaluates the strength of evidence under publication bias when there is only a single study.

**Methods** We first re-interpret t-statistics in a one-tailed hypothesis test in terms of their associated p-values when there is extreme publication bias, that is, when no null findings are published. We then consider the consequences of different degrees of publication bias. We show that under even moderate levels of publication bias adjusting one's p-values to insure Type I error rates of either 0.05 or 0.01 result in far higher t-values than those in a conventional t-statistics table. Under a conservative assumption that publication bias occurs 20 percent of the time, with a one-tailed test at a significance level of 0.05, a t-value equal or greater than 2.311 is needed. For a two-tailed test the appropriate standard would be equal or above 2.766. Both cutoffs are far higher than the traditional ones of 1.645 and 1.96. To achieve a p-value less than 0.01, the adjusted t-values would be 2.865 (one-tail) and 3.254 (two-tail), as opposed to the traditional values 2.326 (one-tail) and 2.576 (two-tail). We illustrate our approach by applying it to evaluate the hypothesis tests in recent issues of *Criminology* and *Journal of Quantitative Criminology* (*JQC*).

**Conclusion** Under publication bias much higher t-values are needed to restore the intended p-value. By comparing the observed test scores with the adjusted critical values, this paper provides a rough rule of thumb for readers to evaluate the degree to which a reported positive result in a single publication reflects a true positive effect. Further measures to increase the reporting of robust null findings are needed to ameliorate the issue of publication bias.

**Keywords** Publication bias · t-value · p-value · Significance test

✉ Christopher Winship
cwinship@wjh.harvard.edu

Xiaolin Zhuo
xiaolinzhuo@fas.harvard.edu

[1] Department of Sociology, Harvard University, Cambridge, USA

## Introduction

A key issue is how to interpret t-statistics when publication bias is present. Publication bias occurs when the results of a study, most commonly the statistical significance of the results, determine whether it is published. This may be due to authors being reluctant to write up or submit null findings (Cooper et al. 1997; Franco et al. 2014); authors may selectively report significant findings after "p-hacking" using "researcher degrees of freedom," that is, actively exploring various analytic options to achieve statistically significant results (Simmons et al. 2011); authors may make, through the "garden of forking paths", data-contingent analytic choices without explicit "p-hacking" (Gelman and Loken 2014);[1] or reviewers and editors may reject studies with statistically insignificant results regardless of their rigorous designs (Franco et al. 2014). Considerable evidence of publication bias exists across many disciplines, including medicine (Begg and Berlin 1988; Easterbrook et al. 1991; Ioannidis et al. 2014), psychology (John et al. 2012; Kühberger et al. 2014; Wicherts et al. 2011), economics (Doucouliagos 2005), political science (Gerber et al. 2001, 2010) and sociology (Gerber and Malhotra 2008).

Evaluating empirical findings under the possible influence of publication bias requires caution. Publication bias increases the likelihood of false positive findings and effect inflation. The results of small, underpowered studies are particularly subject to the selection bias. Their negative findings are more easily dismissed as inconclusive or uninformative, and their positive findings often reflect inflated effect sizes rather than the true effects (Button et al. 2013; Gelman et al. 2017; Ioannidis 2005). As the validity of findings in individual studies becomes compromised, systematic reviews of the literature can no longer balance evidence or uncover the true population effect (Ferguson and Brannick 2012). For example, significant positive effects identified in a pooled analysis of published clinical trials either decrease or disappear after pooling all registered trials (Simes 1986). Inflated effect sizes, due to publication or selection bias, also make original studies harder to reproduce (Open Science Collaboration 2015).

There exist a variety of methods in the literature to detect publication bias and even to recover an unbiased population estimate (Rothstein et al. 2005). These methods typically use a collection of studies. In particular, to evaluate the strength of empirical evidence for a specific effect requires a set of studies that attempt to estimate the same parameter or effect. Such situations, however, are rare in social sciences, especially outside of psychology and economics. Often the support for an empirical finding comes exclusively from a single study. The existing literature offers few guidelines to address publication bias in one individual study.

In this paper, we partially fill this gap first by interpreting t-statistics[2] in a one-tailed hypothesis test in terms of their associated p-values when there is extreme publication bias–only studies with t-values greater than 1.645, that is having a nominal p-value of 0.05, are published. We describe a simple behavioral model of the process of publication bias

[1] A related, but separate problem is to adjust for multiple hypothesis testing (Denton 1985). The False Discovery Rate and other procedures have been developed to correct multiple comparisons (Benjamini and Hochberg 1995; Harvey et al. 2016; Miller 2012). See footnote 5 for further discussion.

[2] In this paper in evaluating the relationship between a t-statistic and p-value we assume that the t-statistics are generated from a normal distribution. This is generally considered valid if one's sample size is infinite, and it is correct to a very close approximation if the degrees of freedom associated with the t-statistic are 50 or greater.

that drives the results. We further show that under such extreme publication bias, the t-values associated with p-values of 0.05 and 0.01 are far higher than those conventionally used in t-tests. Publication bias also changes the interpretation of t-values above but near 1.645 in terms of the evidence they provide for the alternative or the null. Furthermore, using a stricter level of significance or having increased statistical power as a result of a larger sample or larger effect size fails to ameliorate the misalignment between estimated t-values and their conventional p-values.[3] We then consider the consequences of different degrees of publication bias. Importantly, we show that under very moderate levels of publication bias, 20% (that is, a random 20% of null findings are withheld from publication, possibly due to not being written up or submitted, or being rejected by reviewers or editors), fixing one's p-values to insure Type I error rates of either 0.05 or 0.01 results in far higher t-values than those in a conventional t-statistics table. This latter analysis then provides the basis for a set of "rough rules of thumb" for evaluating the strength of empirical evidence under publication bias in a single publication.[4]

Obviously, the best solution for publication bias would be to publish papers with both null and non-null findings. This, undoubtedly, is unrealistic, given the limited journal space and natural priority to publish strong significant findings with potentially greater impact. The set of adjusted critical values we propose serves as an alternative way to evaluate the findings in a single publication in the presence of selection bias. Admittedly our method is imprecise, but is far better than interpreting t-values under publication bias as being equal to the p-values in a traditional table of t-statistics. Furthermore, it is critical to understand that our method is intended to help *readers* assess the relationship between published t-values and their actual associated p-values in determining the level of confidence to place in published findings. *It is explicitly not for authors, reviewers, or editors to use as a criterion of publication*. As we demonstrate below, if authors, reviewers, or editors adopt the larger adjusted critical values and simply become more selective in published results, even higher critical values will be required to restore the intended Type I error rate. We illustrate the use of our approach by applying it to evaluate the hypothesis tests in recent issues of *Criminology* and *Journal of Quantitative Criminology* (*JQC*). Finally, we conclude our paper by discussing the limitations of our method and directions for future research.

## Existing Methods

In this section, we provide a short overview of existing methods for detecting or correcting publication bias. The existing methods are generally for use with a body of multiple studies, which contrasts with our approach that addresses the impact of publication bias on findings reported in a single study. We first discuss common approaches to evaluating publication bias in meta-analysis or when multiple studies estimating the same effect are available. We then describe methods applicable to broader sets of studies.

[3] Of course, a stricter significance level, a larger sample and larger effect size decrease the probability of rejecting the alternative when it is in fact true.

[4] McCrary et al. (2016) developed a model with a similar purpose to ours. Their method can apply to multiple types of hypothesis testing, whereas our approach exclusively focuses on the standard t-test and explores in depth the interpretation of t-values under influence of publication bias. Although McCrary et al. (2016) start from different assumptions about how selection bias works, our models are equivalent under certain parameter specifications.

A common method for detecting publication bias is the funnel plot, which plots effect sizes from individual studies (on the X axis) against a measure of precision, such as sample size or standard error (on the Y axis). Asymmetry in the plot provides likely evidence of publication bias (Peters et al. 2008; Sterne et al. 2005, 2011). Several statistical tests have been developed to measure the asymmetry in funnel plots, such as the Egger test—regressing standardized effect sizes against sampling precision (Egger et al. 1997; Sterne and Egger 2005), rank correlation test—measuring the correlation between standardized effect sizes and sampling variances (Begg and Mazumdar 1994), and the trim and fill method—filling the asymmetric region of the funnel plot with imputed studies and estimating the true effect based on the filled funnel plot (Duval 2005; Duval and Tweedie 2000a, b).

Other methods to assess publication bias in estimates of a single effect that are not built on funnel plots exist. The failsafe N or file-drawer number method calculates the minimum number of additional null results needed to overturn the conclusion drawn from the meta-analysis (Rosenthal 1979). Selection models compare the conventional unweighted mean effect estimate to the estimate weighted by study characteristics such as effect sizes and levels of statistical significance (Hedges and Vevea 2005; Iyengar and Greenhouse 1988).

In addition, scholars have examined publication bias in more general collections of studies. Using frequency distributions, Sterling (1959) first reported the underrepresentation of insignificant results in several psychology journals. Others have analyzed the publication outcomes of a cohort of registered studies (Cooper et al. 1997; Franco et al. 2014). Moreover, if the significance of results determines publication decisions, studies that merely pass the conventional cutoff critical value will, to an appreciable degree not explained by chance, outnumber the studies right below the cutoff. The caliper test measures such discontinuity found at conventional levels of significance to assess the presence of publication bias (Gerber et al. 2010; Gerber and Malhotra 2008). Another approach is the p-curve method that corrects for publication bias using only significant results. The shape of the p-curve, which is the distribution of significant p-values, offers a diagnosis of whether the findings in a general set of studies reflect true effects or merely selective reporting. The p-curve can also be used to estimate the true average effect size in meta-analysis (Simonsohn et al. 2014a, b).

## Single Studies

The aforementioned methods rely on a collection of studies to detect or correct publication bias. The issue remains: how do we evaluate the strength of empirical evidence and adjust for publication bias when only a single study exists? In other words, how do we interpret t-statistics in a single study in light of publication bias? We start by examining the case of extreme bias. Under such bias the actual t-values required to insure the desired Type I error are significantly higher than the conventional critical t-values. In a subsequent section we show that this remains true even when publication bias is modest, for example, 20%.

### Extreme Bias

Consider a model where a potentially unlimited set of researchers independently and sequentially conduct studies, each with new data to determine whether the null hypothesis for a specific effect can be rejected. Publication occurs when the first research result that is significant appears. We restrict our focus to the first time that an effect is published,

ignoring subsequent estimates of the effect of interest or assuming that they do not occur. The censoring of insignificant findings could be done by authors, reviewers or editors. The critical assumption here is that in each analysis the t-statistic under the null hypothesis is drawn from an approximately normal distribution with mean zero and variance one.

A multitude of possible selection processes may be operating to influence publication decisions. As described, our model is applicable to simple publication bias where a null effect is sequentially tested by many researchers on different datasets and a significant effect eventually occurs due to sampling error. Is our model consistent with other sources of selection bias such as p-hacking and garden-path forks of decision making? Answering this question requires evaluating on a model-by-model basis whether the key condition that the t-statistic is drawn from an approximately standard normal distribution is satisfied. As an example, imagine a researcher who keeps re-estimating his or her model including unnecessary control variables that either do not affect the outcome or are uncorrelated with the independent variable whose effect is the interest of the study. In this case, each analysis still draws a new t-statistic from a standard normal distribution. Although such draws may well not be independent, our model does not require an assumption of independence. In contrast, consider a situation where the researcher keeps respecifying the model omitting variables that affect the outcome and are correlated with the key independent variable. In short, they are estimating models where there is an omitted variable problem. In this case, the estimated t-statistic no longer has a mean of zero, and as such our model does not hold. Hopefully, in this type of situation reviewers would point out that the analysis contains serious omitted variable problems.[5]

Under a one-tailed test with the criterion of a p-value of 0.05, assuming the most extreme form of publication bias as defined above entails that only studies that identify an effect greater than 1.645 are published. All insignificant results (effects smaller than 1.645) under such a one-tailed t-test will be put into the so-called file drawer and hidden away. Under this scenario, published effects are effectively being drawn from a normal distribution truncated below at 1.645. Table 1 lists the nominal and actual p-values under the null hypothesis for a range of t-statistics based on this truncated normal distribution. The conventional critical value 1.645 corresponds to a nominal p-value of 0.05, but its actual p-value assuming the null is 1! An actual p-value of 0.05 is in fact associated with a t-statistic of approximately 2.8. In addition, note the actual p-value is 20 times its nominal value. This is because the standard normal distribution after being truncated below 1.645 initially only has an area of 0.05. As a result, when the probability density function is rescaled to achieve an area of one, making it a well-defined density function, all the original nominal p-values are divided by 0.05, equivalent to multiplying by 20.

Next, we consider how much support there is for the null distribution versus the alternative by comparing their relative probability densities. To make our results as clear and simple to understand as possible, we continue to assume extreme publication bias. We consider a null model with mean 0 and standard error of 1 and an alternative model consisting of a normal distribution with mean 3 and standard error of 1. In null hypothesis significance testing, the alternative model or distribution is typically unknown to researchers. Here we specify the alternative distribution N(3, 1) for illustrative purposes, and the intervals we report below depend on this particular choice of the

[5] Although we are not concerned in this paper with the multiple testing problem, our model is consistent with a situation in which a researcher sequentially estimates an unknown number of different null effects and publishes the first effect that is found to be significant.

**Table 1** Test statistics and corresponding p-values under absolute publication bias

| Test statistics | Nominal p-value | Actual p-value | |
|---|---|---|---|
| | | Truncated at 1.645 | Truncated at 2.326 |
| *1.645* | *0.050* | *1.000* | |
| 1.700 | 0.045 | 0.891 | |
| 1.800 | 0.036 | 0.719 | |
| 1.900 | 0.029 | 0.574 | |
| 2.000 | 0.023 | 0.455 | |
| 2.100 | 0.018 | 0.357 | |
| 2.200 | 0.014 | 0.278 | |
| 2.300 | 0.011 | 0.214 | |
| *2.326* | *0.010* | *0.200* | *1.000* |
| 2.400 | 0.008 | 0.164 | 0.820 |
| 2.500 | 0.006 | 0.124 | 0.621 |
| 2.600 | 0.005 | 0.093 | 0.466 |
| 2.700 | 0.003 | 0.069 | 0.347 |
| 2.800 | 0.003 | 0.051 | 0.256 |
| *2.807* | *0.003* | *0.050* | *0.250* |
| 2.900 | 0.002 | 0.037 | 0.187 |
| 3.000 | 0.001 | 0.027 | 0.135 |
| 3.100 | 0.001 | 0.019 | 0.097 |
| 3.200 | 0.001 | 0.014 | 0.069 |
| 3.300 | 0.000 | 0.010 | 0.048 |
| 3.400 | 0.000 | 0.007 | 0.034 |
| 3.500 | 0.000 | 0.005 | 0.023 |
| 3.600 | 0.000 | 0.003 | 0.016 |
| 3.700 | 0.000 | 0.002 | 0.011 |
| *3.719* | *0.000* | *0.002* | *0.010* |
| 3.800 | 0.000 | 0.001 | 0.007 |
| 3.900 | 0.000 | 0.001 | 0.005 |
| 4.000 | 0.000 | 0.001 | 0.003 |
| 4.100 | 0.000 | 0.000 | 0.002 |
| 4.200 | 0.000 | 0.000 | 0.001 |

Rows that correspond to nominal or actual p-values of 0.05 or 0.01 are italicized

alternative model. Assuming a different alternative distribution would produce different numbers, but qualitatively similar results.

Both the null and alternative distributions are truncated at 1.645, a t-value corresponding to a nominal p-value of 0.05. We also assume that null effects are likely to be published, specifically so that among published studies, 50% come from the truncated N(0,1) distribution and 50% from the truncated N(3,1) distribution. Different assumptions about the mixture weights of the null and the alternative models would lead to qualitatively similar results.

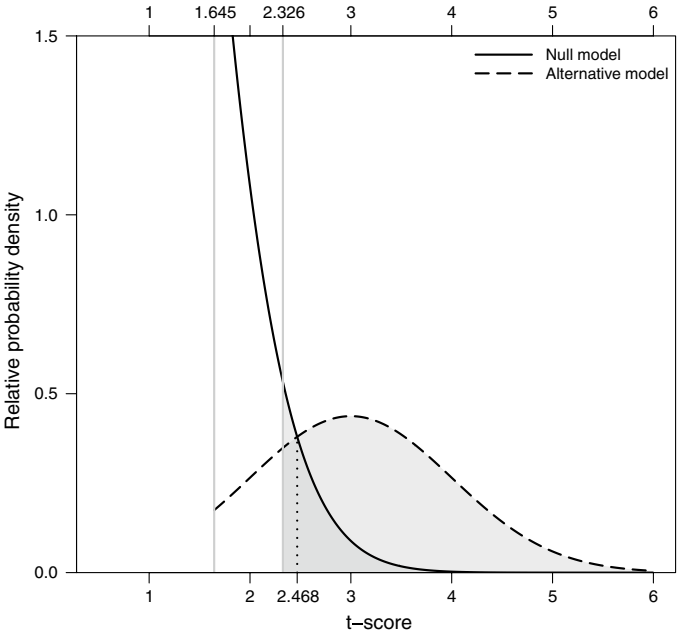


**Fig. 1** Relative probability density functions of test statistics under absolute publication bias. We start with two normal distributions N(0, 1) and N(3, 1) with the first representing the distribution under the null and the second under the alternative. They are scaled so that approximately 95% of the publications come from the null distribution and 5% come from the alternative. Now truncate the two normal distributions at 1.645, the cutoff point for a one-tail test for a p-value of 0.05 (in this figure). In essence, we erase the portion of each distribution that is to the left of the line marked at 1.645. The solid black line represents the null model truncated at 1.645. The dashed black line represents the alternative model truncated at 1.645. The two remaining truncated distributions have equal areas. Thus, a t-statistic that is greater than 1.645 is equally likely to have come from either the truncated null distribution or the truncated alternative distribution, which is consistent with our initial setup. The two density curves cross at 2.468 (marked on the X-axis). t-values within the interval (1.645, 2.468) reflect greater support for the null model than the alternative hypothesis. When a stricter significance level of 0.01 is adopted, we now truncate the original normal distributions at 2.326 (the conventional t-value associated with a p-value of 0.01). Here we are erasing the portion of each distribution that is to the left of the gray line marked at 2.326. The area (shaded) under the null is now approximately 0.2 and under the alternative is 0.82. The range of t-values over which the null model is more likely than the alternative then becomes (2.326, 2.468)

Under complete publication bias only statistically significant results are published. Under this setup, the t-statistics for published effects (observed by readers) effectively follow an even mixture of a N(0, 1) and N(3, 1) distribution, both truncated at 1.645 (see Fig. 1).

A number of properties of the new distributions are noteworthy. Although only "statistically significant" estimates are observed (only values greater than the conventional critical value 1.645), the alternative model is not appreciably more likely than the null model across a range of t-values greater than 1.645. In fact, as shown in Fig. 1 t-values between 1.645 and 2.468 are more likely to be observed under the null. In other words, a relatively large t-value between 1.645 and 2.468, which provides sufficient evidence to reject the null under the traditional unbiased model, actually provides greater support for the null than alternative under our scenario.

The fact that a t-value above some critical cutoff value may provide greater support for the null than some specific alternative is also true in the absence of publication bias. Since the goal of this paper is to make readers aware of the problems in interpreting t-statistics we believe it is important to point out that t-statistics above some cutoff value do not necessarily mean that the data provide stronger support for a particular alternative than the null. Furthermore, given the problems publication bias creates, we assume that some readers will wonder if a stricter significance level for publication or a larger sample size will alleviate the problems we have identified. In the next two sections we show that neither ameliorates the problems with interpreting t-values above a conventional cutoff as providing greater support for the alternative than the null.

### Stricter Significance Level

Can the above problems be ameliorated if authors, editors and reviewers adopt a stricter significance level, say a level of 0.01? We start by examining the consequences of a stricter significance level for the interpretation of t-values in terms of their associated p-values.[6] After restricting the publication criterion to the significance level of 0.01, the published test statistics now follow a standard normal distribution truncated at 2.326. As shown in the last column in Table 1, a t-statistic equal to 2.326 has a nominal p-value of 0.01, but under publication bias with the new stricter level of significance it has an actual p-value of 1! This is in contrast to when publication bias occurred at 0.05, where a t-statistic of 2.326 had an actual p-value of 0.2. The t-statistic associated with a true p-value of 0.01 becomes 3.719, larger than 3.3 when the 0.05 level of significance is used as the criterion for publication. The effects of selection bias are now far worse! A moment of thought, however, should indicate that this is not surprising—a stricter cutoff is equivalent to more severe selection bias. Lower cutoff values would improve the situation. Nondiscriminatory publication of both null and non-null findings would solve it, but as already noted this is unrealistic. The consequences of adopting a stricter significance level further corroborate our emphasis that our approach is not intended for authors, reviewers, and editors to use as a new standard of publication. Their doing so only makes the problems associated with publication bias far worse.

Second, we can examine the change in the range of t-values that reflect greater support for the null versus the alternative following the use of a stricter significance level. Back to our hypothetical setting with a null model with mean 0 and an alternative model with mean 3, which we assume here are equally likely and are both truncated at 1.645. When only results significant at the 0.01 level are published, the range in which the null model is more probable than the alternative has shifted from (1.645, 2.468) to (2.326, 2.468) (Fig. 1). The range over which the null model is more probable than the alternative is now narrower, lowering the chance of a type I error. This, however, is achieved at the cost of excluding all t-values below 2.326 in both the null and the alternative models, hence increasing the

[6] As in the traditional context of significance testing, a higher statistical level for rejection of the null will tautologically lead to a lower Type I error. At the same time, it will increase the probability of a Type II error. Furthermore, as the level of significance increases, the portion of positive results that are in fact truly nonnull, what is known at the Predicted Positive Probability or PPV, will also increase. Although adopting a stricter significance level improves PPV, our question is whether the problems with interpreting t-values with regards to p-values are ameliorated at all.

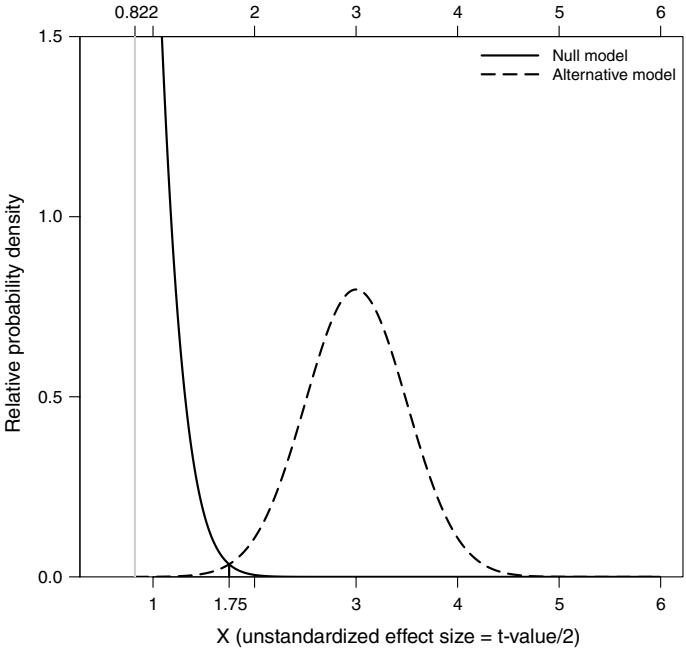


**Fig. 2** Relative probability density functions of unstandardized effect sizes under absolute publication bias when sample size increases by four times. The solid black line represents the probability density function (PDF) of the null model, the normal distribution with mean 0 and standard error 0.5 truncated at 0.822 (marked by a grey vertical line). The dashed black line represents the PDF of the alternative model, the normal distribution with mean 3 and standard error 0.5 truncated at 0.822. As the sample size increases by four times, the distribution of the unstandardized effect sizes (as shown here) shrinks towards its mean. The two relative density curves cross at 1.75 (marked on the X-axis). Unstandardized effect sizes from 0.822 to 1.75, equivalent to t-values from 1.645 to 3.5, reflect greater support for the null model than the alternative

chance of a type II error. Furthermore, it remains true that large t-values up to 2.468 still offer greater support for the null than the alternative.

If we raise the significance level to a more extreme value, say 0.001, the conventional critical value associated is 3.09, exceeding 2.468 the point at which the two truncated normal density curves cross, and we will no longer observe any interval over which the null is more likely than the alternative. This improvement, however, is again inescapable from the cost of dropping all studies with t-values below 3.09 and thus, increasing the risk of a Type II error.

## Larger Sample Size

What are the consequences for interpreting t-values under publication bias when the study sample size increases?[7] In terms of interpretation of t-values under the null nothing changes

[7] It is well known that a larger sample size will increase the power of a study, as well as the Positive Predicted Value (PPV). However, we are considering a separate issue of interpreting t-statistics with regards to p-values following a change in study sample size.

from our previous analysis. The results in Table 1 hold independent of sample size. However, the degree to which a t-value provides evidence for the null versus the alternative can alter radically. We illustrate this again using the comparison of the null model with mean 0 and the alternative model with mean 3. A parallel analysis of the consequences of having a larger effect would produce qualitatively similar results to those below.

Assume that we increase sample sizes by a factor of four, resulting in a standard error of 0.5 for both the null and alternative distributions. Figure 2 (*unstandardized* effect sizes plotted on the same scale as in Fig. 1) shows that since the standard deviation of both distributions has decreased by half, the two distributions have shrunk towards their respective means. The effect associated with a nominal p-value 0.05 becomes $1.645/2 = 0.822$, the truncation point for both the null and the alternative models. Except for this rescaling, the relative density function has not changed with respect to the null—the original normal distribution has only an area of 0.05 to the right of the truncation point. For the alternative, though, almost the entire distribution is to the right of the truncation point. The range of effect sizes above the truncation point of 0.822 that represent greater support for the null is now quite a bit larger, going from 0.822 to 1.75, which correspond to t-values from 1.645 to 3.5. What is happening here is that the increased sample size has caused the alternative distribution to shrink towards its mean, thus resulting in it pulling away from the null and expanding the range of t-values that are more likely under the null model than under the alternative [(1.645, 3.5) vs. (1.645, 2.468) previously]. To see more clearly why this result makes sense, suppose that the alternative had mean 10 and standard deviation 1. Then t-values of say even 4, would be farther away from the mean of the alternative model (6 standard deviations) than from the mean of the null (4 standard deviations) and extremely unlikely under the alternative hypothesis.

## Adjustment to Critical Values

We have demonstrated that conventional critical values in the standard t-test are no longer informative under absolute publication bias. Specifically, the conventional critical value is no longer a correct threshold to maintain the level of significance at 0.05 (or 0.01). It also fails to provide stronger evidence against the null even with a stricter significance level or a larger sample size. In this section, building on our findings about the unbiased model and the absolute publication bias model, we offer *readers* (not authors, reviewers, or editors) guidelines to evaluate the strength of evidence in published works under publication bias.

The scenario of absolute publication bias, discussed above, in which no insignificant results are published at all, may be unrealistic. Authors, reviewers, and editors may commit publication bias some of the time, but not at other times. If this is the case, the parameter of interest under possible influence of publication bias follows a mixture model of a normal distribution and a truncated normal distribution, both with mean zero and variance one. Let $F$ be the cumulative distribution function of the standard normal distribution and $F_{Trunc}$ be the cumulative distribution function of the truncated normal distribution with mean zero and variance one, cut off below the conventional critical value corresponding to a widely accepted significance level, which is commonly 0.05. Let D be the indicator variable of whether a result is published. The cumulative distribution function of the *published* test statistic T is the following conditional probability:

$$G(T|D = 1) = (1 - p) \cdot F(T) + p \cdot F_{Trunc}(T).$$

**Table 2** Adjusted critical values under publication bias

| Severity of publication bias | One-tailed test | | Two-tailed test | |
|---|---|---|---|---|
| | $p=0.05$ | $p=0.01$ | $p=0.05$ | $p=0.01$ |
| 0.00 (no publication bias) | 1.645 | 2.326 | 1.960 | 2.576 |
| 0.01 | 1.728 | 2.391 | 2.097 | 2.688 |
| 0.05 | 1.949 | 2.567 | 2.388 | 2.930 |
| 0.10 | 2.114 | 2.702 | 2.569 | 3.084 |
| 0.15 | 2.227 | 2.795 | 2.683 | 3.183 |
| 0.20 | 2.311 | 2.865 | 2.766 | 3.254 |
| 0.25 | 2.378 | 2.922 | 2.830 | 3.311 |
| 0.30 | 2.434 | 2.969 | 2.883 | 3.357 |
| 0.35 | 2.482 | 3.010 | 2.928 | 3.396 |
| 0.40 | 2.523 | 3.045 | 2.967 | 3.431 |
| 0.45 | 2.560 | 3.077 | 3.000 | 3.461 |
| 0.50 | 2.593 | 3.105 | 3.031 | 3.487 |
| 0.55 | 2.622 | 3.130 | 3.058 | 3.512 |
| 0.60 | 2.649 | 3.154 | 3.083 | 3.534 |
| 0.65 | 2.674 | 3.175 | 3.106 | 3.554 |
| 0.70 | 2.697 | 3.195 | 3.127 | 3.573 |
| 0.75 | 2.719 | 3.213 | 3.146 | 3.590 |
| 0.80 | 2.738 | 3.231 | 3.165 | 3.606 |
| 0.85 | 2.757 | 3.247 | 3.182 | 3.622 |
| 0.90 | 2.775 | 3.262 | 3.198 | 3.636 |
| 0.95 | 2.791 | 3.277 | 3.213 | 3.649 |
| 1.00 (absolute publication bias) | 2.807 | 3.291 | 3.227 | 3.662 |

Adjusted critical values for one-tailed tests are calculated with the assumption that the truncated normal component of the mixture model is cut off below 1.645, the conventional critical value corresponding to the significance level of 0.05 in one-tailed testing. The cutoff value used in two-tailed tests is 1.96, the conventional critical value corresponding to the significance level of 0.05 in two-tailed testing

The mixture weights, $(1 - p)$ and $p$, indicate the level of selection bias. They represent, respectively, the probability an observed test statistic comes from the normal distribution (absence of publication bias) and the truncated normal distribution (the most severe form of publication bias). The actual mixture weights are unknown and are most likely to vary considerably by research situation. Thus, we present the adjusted critical values associated with a wide range of degrees of publication bias (see Table 2). We assume that the truncated normal component of the mixture model is cut off below the conventional critical value for the common 0.05 level of significance, that is, 1.645 for one-tailed testing and 1.96 for two-tailed testing.

When publication bias is absent, the conventional critical values hold. A critical value of 1.645 corresponds to the significance level of 0.05 in one-tailed testing, and 1.96 corresponds to the significance level of 0.05 in two-tailed testing. As publication bias worsens (the truncated normal distribution receives more weight in the mixture model), the critical value needed to restore the intended significance level steadily rises until we reach the extreme point when no insignificant results are published. Under absolute publication bias, only test statistics equal to or greater than 2.807 can insure a false positive rate lower than

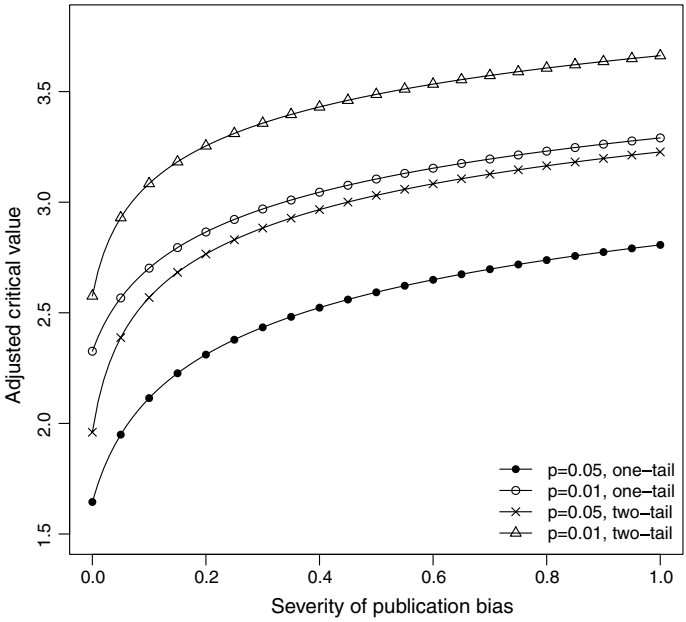


**Fig. 3** Adjusted critical values. Adjusted crucial values are calculated based on the mixture model of a normal distribution and a truncated normal distribution, both with mean one and variance zero. The mixture weight of the truncated normal component is represented by the severity of publication bias on the X-axis. In one-tailed testing, the truncated normal distribution is cut off below 1.645, the conventional critical value corresponding to the significance level 0.05. In two-tailed testing, the truncated normal distribution is cut off below 1.96, the conventional critical value corresponding to the significance level 0.05

5% in a one-tailed test. Two-tailed testing requires a threshold of 3.227. These adjusted critical values offer a benchmark when examining the degree to which a published positive effect indeed reflects a true positive population effect. Table 2 also presents adjusted critical values at varying degrees of publication bias when a significance level of 0.01 is desired.

Figure 3 plots adjusted critical values against changing levels of publication bias for both one-tailed testing and two-tailed testing and for both 0.05 and 0.01 levels of significance. Similar to standard null hypothesis statistical testing, two-tailed testing requires higher critical values than one-tailed testing to achieve the same level of significance. Moreover, we observe that although critical values monotonically increase as publication bias worsens, the rate of increase in fact declines. The greater rate of increase at the lower level of publication bias implies that even a relatively small degree of publication bias can significantly inflate the actual p-value and demand a sizable increase in critical values to restore the intended Type I error rate.[8] For instance, if publication bias happens 20% of the time, the critical value to maintain the significance level 0.05 in a one-tailed test is raised from 1.645 to 2.311 or increases by a magnitude of 0.666. The difference between adjusted critical values with no publication bias and with 100% publication bias is 1.162. This means that more than half (0.666/1.162 = 57%) of the total change in the adjusted

[8] We thank one of the reviewers of the original version of this paper for pointing out this important fact.

critical value from the absence of publication bias to absolute publication bias occurs when publication bias is only 20%. This finding highlights that we should treat publication bias seriously even if only a small amount of publication bias exists in a field.

Often, if not most of the time, when trying to evaluate the results of a single study it may be impossible to estimate the degree of publication bias due to possible influence of p-hacking, multiple testing, quality of research design, etc. Cooper et al. (1997) documented the bias against null findings by showing that among a set of IRB-approved and completed studies, 74% of significant results were submitted for publication, whereas only 4% of nonsignificant results were submitted. Following the outcomes of a cohort of registered high-quality experiments, Franco et al. (2014) found that only 20% of null results were published, whereas about half of mixed results (only some hypotheses are supported by statistical tests) and three fifths of strong results (all or most hypotheses are supported) were published. In fact, 65% of null results were never written up.

We would thus argue that publication bias occurs only 20% of the time is a conservative assumption. Under this conservative assumption, if we want to do a one-tailed test at a significance level of 0.05, then roughly a t-value equal or greater than 2.311 is needed. For a two-tailed test the appropriate standard would be equal or above 2.766. Both cutoffs are far higher than the traditional ones, 1.645 and 1.96. To achieve a p-value less than 0.01, the respective adjusted t-values would be 2.865 (one-tail) and 3.254 (two-tail), as opposed to the traditional values 2.326 (one-tail) and 2.576 (two-tail).

## Examples

In this section, we apply adjusted critical values to evaluate results of standard t-tests in the studies published in the 2016 issues of *Criminology* and *Journal of Quantitative Criminology* (*JQC*). A search for empirical quantitative papers that run t-tests and report t-values or information needed to construct t-values (for example, coefficients and standard errors or confidence intervals) yielded 19 articles in *Criminology* and 20 articles in *JQC* respectively. Within these articles, we further restricted our analysis to key variables that are associated with hypothesis tests or discussed extensively in the main text. Two coders (one of the authors and a research assistant) independently collected test scores from the qualified articles. The coders then compared their datasets and resolved discrepancies to create the final dataset. Our final sample consists of 979 t-scores extracted from *Criminology* articles published in 2016 (with a median of 36 t-scores per article) and 1115 from *JQC* (with a median of 34).[9]

Next, we compare the distribution of t-statistics to the conventional critical value (1.96) for two-tailed tests at a significance level of 0.05 and the adjusted critical value (2.766) under the conservative assumption that publication bias occurs 20% of the time.[10] The distributions of t-test statistics in the two journals display similar patterns (see Fig. 4). They are both positively skewed. A substantial portion of observed t-values falls under the

[9] We excluded effect estimates with standard errors that were not reported or reported equal to zero.

[10] We applied the criteria for two-tailed tests because two-tailed tests are commonly carried out in the *Criminology* and *JQC* publications we examined, even when the hypotheses imply that one-tailed tests are needed. The widespread adoption of two-tailed tests is partly attributable to the default setting in statistical software and the more conservative nature of two-tailed tests. Further analysis is needed to document this inconsistency between hypotheses and choice of statistical tests.

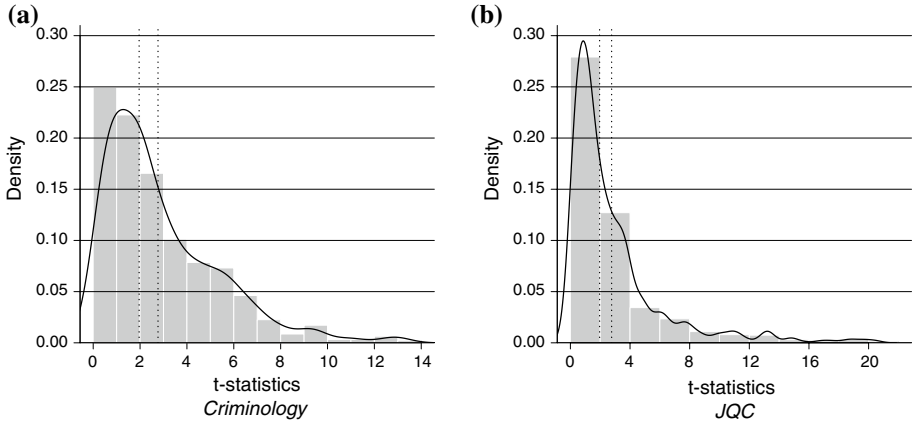


**Fig. 4** Distributions of t-statistics in Selected Criminology Journals. The t-statistics were collected from studies published in the *Criminology* and *JQC* 2016 issues. Only primary independent variables that are linked to hypotheses or discussed extensively in the main text were included. The t-statistics were either directly retrieved from the tables in the publications or calculated using coefficients, standard errors, odds ratios, or confidence intervals. For visualization purpose only, the distribution of t-scores in *Criminology* publications is cut off above the 95th percentile, and the distribution of t-scores in *JQC* is cut off above the 90th percentile. The histogram of the distribution is overlaid with a density plot. The dotted vertical lines represent the conventional critical value 1.96 and the adjusted critical value 2.766. **a** Criminology, **b** Journal of Quantitative Criminology

conventional cutoff value 1.96 (42% in *Criminology* and 49% in *JQC*). Around 40% of t-statistics in both journals surpassed the adjusted critical value (2.766) under the conservative assumption of 20% publication bias. The remaining 15% of t-scores in *Criminology* and 11% in *JQC* fall into the grey area between the conventional and the adjusted critical values. If we consider nominally "statistically significant" effect estimates alone (those surpassing the conventional threshold 1.96), then 26% of t-scores in *Criminology* and 22% in *JQC* lie between the conventional cutoff value and the adjusted one. This implies that about a quarter of the published positive findings may in fact be support for the null hypothesis rather than the alternative, under the conservative assumption that a modest degree (20%) of publication bias exists. As such, the conclusions from these published results may not be warranted.

It is important to recognize that the sizable portion of insignificant findings published in *Criminology* and *JQC* does not offer conclusive evidence for either the prevalence or rareness of publication bias in the field of criminology. Without a representative sample of published and unpublished papers, it is not possible to estimate the degree of selection. We note that multiple testing is a common practice in the articles we examined,[11] and that insignificant test statistics always co-exist with other significant ones in each publication. No article in our sample was published with only insignificant findings. This observation is consistent with Franco et al.'s (2014) result that a mixed set of findings are highly likely to be published. The over-emphasis on statistical significance and the multiple comparisons problem both contribute to the rise of mixed-finding studies. As Gelman and his

[11] We could identify only two papers (Wolfe et al. 2016; Wooditch and Weisburd 2016) that considered the multiple testing problem.

co-authors note, "That there are a large number of reported insignificant effect estimates in the literature is not contrary to a potential significance filter" (Gelman et al. 2017: 6). What the above empirical analysis does show is that at least some nonsignificant findings do get published.

Drawing strong inferences about the level of publication bias in criminology using our method of adjusted critical values is not appropriate as our method specifically applies to addressing publication bias in a single publication. As discussed in the previous section on existing methods, a number of advanced statistical methods are already available to assess the degree of publication bias in meta-analysis of multiple studies estimating the same effect or in a general collection of studies. These methods are better positioned to assess the level of publication bias in a field or compare across journals or disciplines. Meta-analyses in criminology, for instance, have found mixed evidence of the presence of publication bias (Piquero et al. 2009; Rothstein 2008).

## Conclusion

In this paper, we have presented a new approach to interpreting t-values and evaluating the strength of evidence in published research under publication bias. Contrary to existing methods that utilize collections of studies, we assess the impact of publication bias when there is only one published study. We show that when significant results are more likely to get published, the critical values needed to insure the desired significance level, or Type I error rate, are in fact quite a bit higher than the conventional critical values. More specifically, we model, under publication bias, the distribution of the parameter of interest, assuming the null model is true, as a mixture of a normal distribution and a truncated normal distribution. Based on this mixture model, we further propose an adjustment to the critical values to restore the intended Type I error rate. With a conservative assumption that publication bias happens 20% of the time, a one-tailed test at a significance level of 0.05 implies a critical value of 2.311. For a two-tailed test the appropriate standard would be equal to or above 2.766. Both cutoffs far exceed the conventional critical values, 1.645 and 1.96 respectively. To maintain a p-value less than 0.01, the respective adjusted t-values would be 2.865 (one-tail) and 3.254 (two-tail), as opposed to the traditional values 2.326 (one-tail) and 2.576 (two-tail). If a reader believes that the degree of publication bias exceeds 20%, Table 2 lists appropriate t-values for assessing whether an effect is significant at the 0.05 or 0.01 level for a wide range of levels of publication bias.

Comparing the observed test scores with the adjusted critical values provides a rough rule of thumb for *readers* to evaluate the degree to which a reported positive result in a single publication reflects a true positive effect. We emphasize that our approach is intended for *readers* only as a standard of evaluation of evidence in individual published studies. It is not intended to use as a standard of publication by authors, reviewers, or editors. If authors, reviewers, or editors were to adopt more stringent significance test criteria, publication bias would only worsen.

It is important to point out that even though our adjustment to critical values might seem equivalent to raising the level of significance and adopting a higher critical value, they are different in that the latter is associated with a stricter level of significance, whereas our approach is used to restore the intended level of significance in response to its inflation under the presence of publication bias. In a recently published paper, Benjamin et al. (2018) suggest adopting a p-value of 0.005, equivalent to a traditional t-value of 2.576

(one-tail) as indicating strong evidence for rejecting the null. Within our framework using a true p-value of 0.05 in a one-tailed test, this would be equivalent to assuming that there was almost a 50% rate of publication bias.

In this paper, we have developed a simple model for dealing with publication bias in a single study. It offers a useful starting point to evaluate the strength of evidence in findings reported in a single publication, complementing existing methods that are used with collections of studies. Our simple model, however, is not intended to nor is it able to fully account for many crucial features in the real publication process, such as the multiple comparison problem, differential probability of publication based on the significance level or power, mismatch of two-tailed testing with directional hypotheses, or various p-hacking or "garden of forking paths" practices that result in false positive findings. Moreover, we exclusively focused on the standard t-test and discussed in depth the implication of publication bias for this particular type of hypothesis testing. Although the standard t-test is arguably the most popular hypothesis test in social sciences, there exist other kinds of hypothesis tests, and more general approaches, such as the model developed by McCrary et al. (2016), are needed to evaluate publication bias in the other types of hypothesis tests.

Given that with any particular published study it is difficult to know the degree of publication bias that it survived, we argue that it is not appropriate to adopt any one specific t-value as a precise value for deciding whether or not to reject the null hypothesis. Rather the take-away from this paper should be that much higher t-values are needed for assessing the strength of evidence that exists for the null (a consistent recommendation made by others, albeit backed by different rationales, such as in Benjamin et al. (2018) and Harvey et al. (2016)). Table 2 allows for this assessment by providing t-values for one- and two-tailed t-tests at p-values of 0.05 and 0.01 under various degrees of publication bias. These can be thought of as creating "rough rules of thumb" for what constitutes support for the null under different assumptions.

In addition to using our approach, what might be done by authors, reviewers, and editor to help ameliorate publication bias? Null findings constitute a valuable part of scholarly knowledge. Efforts that would help would be to take direct measures to increase the reporting of robust null findings, for example, requiring a registry of research proposals (Nosek and Lakens 2014), sharing datasets from published studies (Alsheikh-Ali et al. 2011; Wicherts et al. 2011), enforcing a two-stage review process (the first stage for research design and the second stage for implementation and results) (Greve et al. 2013; Smulders 2013), providing high-status publication outlets for null findings, and mandating the publication of all funded studies (Franco et al. 2014). We should also encourage replications of published studies to strengthen cumulative empirical evidence (Koole and Lakens 2012).[12]

**Acknowledgements** We would like to thank the editors and anonymous reviewers for their most helpful comments on earlier drafts of this paper. We would also like to thank Stephanie Wu for her research assistance.

[12] The codes that are used to generate tables and figures in the text are available at https://github.com/xiaolinzhuo/pub_bias.

## References


Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA (2011) Public availability of published research data in high-impact journals. PLoS ONE 6(9):e24357

Begg CB, Berlin JA (1988) Publication bias: a problem in interpreting medical data. J R Stat Soc Ser A (Stat Soc) 151(3):419–463

Begg CB, Mazumdar M (1994) Operating characteristics of a rank correlation test for publication bias. Biometrics 50(4):1088–1101

Benjamin DJ et al (2018) Redefine statistical significance. Nat Hum Behav 2:6–10

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol) 57(1):289–300

Button KS et al (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14(5):365–376

Cooper H, DeNeve K, Charlton K (1997) Finding the missing science: the fate of studies submitted for review by a human subjects committee. Psychol Methods 2(4):447–452

Denton FT (1985) Data mining as an industry. Rev Econ Stat 67(1):124–127

Doucouliagos C (2005) Publication bias in the economic freedom and economic growth literature. J Econ Surv 19(3):367–387

Duval S (2005) The trim and fill method. In: Rothstein HR, Sutton AJ, Borenstein M (eds) Publication bias in meta-analysis: prevention, assessment and adjustments. Wiley, New York, pp 127–144

Duval S, Tweedie R (2000a) A nonparametric 'trim and fill' method of accounting for publication bias in meta-analysis. J Am Stat Assoc 95(449):89–98

Duval S, Tweedie R (2000b) Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 56(2):455–463

Easterbrook PJ, Gopalan R, Berlin JA, Matthews DR (1991) Publication bias in clinical research. Lancet 337(8746):867–872

Egger M, Smith GD, Schneider M, Minder C (1997) Bias in meta-analysis detected by a simple, graphical test. BMJ 315(7109):629–634

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

Franco A, Malhotra N, Simonovits G (2014) Publication bias in the social sciences: unlocking the file drawer. Science 345(6203):1502–1505

Gelman A, Loken E (2014) The statistical crisis in science. Am Sci 102(6):460–465

Gelman A, Skardhamar T, Aaltonen M (2017) Type M error might explain Weisburd's Paradox. Retrieved 7 Oct 2017. http://www.stat.columbia.edu/~gelman/research/published/weisburd_28.05.2017.pdf

Gerber AS, Malhotra N (2008) Publication bias in empirical sociological research: do arbitrary significance levels distort published results? Sociol Methods Res 37(1):3–30

Gerber AS, Green DP, Nickerson D (2001) Testing for publication bias in political science. Political Analysis 9(4):385–392

Gerber AS, Malhotra N, Dowling CM, Doherty D (2010) Publication bias in two political behavior literatures. Am Polit Res 38(4):591–613

Greve W, Bröder A, Erdfelder E (2013) Result-blind peer reviews and editorial decisions: a missing pillar of scientific culture. Eur Psychol 18(4):286–294

Harvey CR, Liu Y, Zhu H (2016) … and the cross-section of expected returns. Rev Financ Stud 29(1):5–68

Hedges LV, Vevea J (2005) Selection method approaches. In: Rothstein HR, Sutton AJ, Borenstein M (eds) Publication bias in meta-analysis: prevention, assessment and adjustments. Wiley, New York, pp 145–174

Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8):e124

Ioannidis JPA, Munafò MR, Fusar-Poli P, Nosek BA, David SP (2014) Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends Cogn Sci 18(5):235–241

Iyengar S, Greenhouse JB (1988) Selection models and the file drawer problem. Stat Sci 3(1):109–117

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

Koole SL, Lakens D (2012) Rewarding replications: a sure and simple way to improve psychological science. Perspect Psychol Sci 7(6):608–614

Kühberger A, Fritz A, Scherndl T (2014) Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. PLoS ONE 9(9):e105825

McCrary J, Christensen G, Fanelli D (2016) Conservative tests under satisficing models of publication bias. PLoS ONE 11(2):e0149590

Miller RG Jr (2012) Simultaneous statistical inference. Springer, Berlin

Nosek BA, Lakens D (2014) Registered reports: a method to increase the credibility of published results. Soc Psychol 45(3):137–141

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349(6251):aac4716

Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L (2008) Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. J Clin Epidemiol 61(10):991–996

Piquero AR, Farrington DP, Welsh BC, Tremblay R, Jennings WG (2009) Effects of early family/parent training programs on antisocial behavior and delinquency. J Exp Criminol 5(2):83–120

Rosenthal R (1979) The file drawer problem and tolerance for null results. Psychol Bull 86(3):638–641

Rothstein HR (2008) Publication bias as a threat to the validity of meta-analytic results. J Exp Criminol 4(1):61–81

Rothstein HR, Sutton AJ, Borenstein M (eds) (2005) Publication bias in meta-analysis: prevention, assessment and adjustments. Wiley, New York

Simes RJ (1986) Publication bias: the case for an international registry of clinical trials. J Clin Oncol 4(10):1529–1541

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

Simonsohn U, Nelson LD, Simmons JP (2014a) P-curve: a key to the file-drawer. J Exp Psychol Gen 143(2):534–547

Simonsohn U, Nelson LD, Simmons JP (2014b) P-curve and effect size: correcting for publication bias using only significant results. Perspect Psychol Sci 9(6):666–681

Smulders YM (2013) A two-step manuscript submission process can reduce publication bias. J Clin Epidemiol 66(9):946–947

Sterling TD (1959) Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. J Am Stat Assoc 54(285):30–34

Sterne JAC, Egger M (2005) Regression methods to detect publication and other bias in meta-analysis. In: Rothstein HR, Sutton AJ, Borenstein M (eds) Publication bias in meta-analysis: prevention, assessment and adjustments. Wiley, New York, pp 99–110

Sterne JAC, Becker BJ, Egger M (2005) The funnel plot. In: Rothstein HR, Sutton AJ, Borenstein M (eds) Publication bias in meta-analysis: prevention, assessment and adjustments. Wiley, New York, pp 75–98

Sterne JAC et al (2011) Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 343:d4002

Wicherts JM, Bakker M, Molenaar D (2011) Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. PLoS ONE 6(11):e26828

Wolfe SE, Nix J, Kaminski R, Rojek J (2016) Is the effect of procedural justice on police legitimacy invariant? Testing the generality of procedural justice and competing antecedents of legitimacy. J Quant Criminol 32(2):253–282

Wooditch A, Weisburd D (2016) Using space-time analysis to evaluate criminal justice programs: an application to stop-question-frisk practices. J Quant Criminol 32(2):191–213