The Replication Paradox: Combining Studies Can Decrease Accuracy of Effect Size Estimates

Michèle B. Nuijten,

Marcel A. L. M. van Assen,

Coosje L. S. Veldkamp,

and

Jelte M. Wicherts

Tilburg University

Author Note

Correspondence concerning this article should be addressed to Michèle Nuijten, PO Box 90153, 5000 LE Tilburg, The Netherlands, M.B.Nuijten@tilburguniversity.edu.

The preparation of this article was supported by VIDI grant number 016-125-385 from the Netherlands Organization for Scientific Research (NWO).

Abstract

Replication is often viewed as the demarcation between science and non-science. However, contrary to the commonly held view, we show that in the current (selective) publication system replications may increase bias in effect size estimates. Specifically, we examine the effect of replication on bias in estimated population effect size as a function of publication bias and the studies' sample size or power. We analytically show that incorporating the results of published replication studies will in general not lead to less bias in the estimated population effect size. We therefore conclude that mere replication will not solve the problem of overestimation of effect sizes. We will discuss the implications of our findings for interpreting results of published and unpublished studies, and for conducting and interpreting results of meta-analyses. We also discuss solutions for the problem of overestimation of effect sizes, such as discarding and not publishing small studies with low power, and implementing practices that completely eliminate publication bias (e.g., study registration).

Keywords: replication, effect size, publication bias, power, meta-analysis

Imagine that you want to estimate the effect size of a certain treatment. To this end, you search for articles published in scientific journals and you come across two articles that include an estimation of the treatment effect. The two studies can be considered exact replications because the population, designs and procedures of the included studies are identical. The only difference between the two studies concerns their sample size: one study is based on 40 observations (a small study; S), whereas the other study is based on 70 observations (a larger study; L). The following questions are now relevant: How do you evaluate this information? Which effects would you include to get the most accurate estimate of the population effect? Would you evaluate only the small study, only the large study, or both? And what if you would have come across two small or two large studies?

To get an idea about the intuitions researchers have about these questions, we administered a short questionnaire (see Appendix 1) among three groups of subjects, with supposedly different levels of statistical knowledge: second year's psychology students (N=106; paper survey administered during statistics tutorials; Dutch translation), social scientists (N=360; online survey), and quantitative psychologists (N=31; paper survey administered at the 78th Annual Meeting of the Psychometric Society). In the questionnaire we presented different hypothetical situations with combinations of small and large studies, all published in peer-reviewed journals, and asked which situation would yield the most accurate estimate of the effect of the treatment in the population. Accuracy was described in the questionnaire as "the closeness of the estimate to the population effect, inversely related to the bias of an estimate". We list the different situations and responses in Table 1.[1]

Table 1.

Results of the questionnaire to assess researchers' intuitions about the value of replication. Answers of 106 psychology students (PS), 360 social scientists (SS), and 31 quantitative psychologists (QP). S = Small published study with 40 observations; L = Large published study with 70 observations.

	“Which situation (A or B) yields the most accurate estimate of the effect of the treatment in the population?”		**Proportion of subsample that endorses the answer category**								
	Situation A	Situation B	Situation A more accurate			Situation B more accurate			Situation A and B equally accurate		
			PS	SS	QP	PS	SS	QP	PS	SS	QP
Question 1	L*	S	**.972**	**.857**	**.871**	.019	.036	.032	.009	.108	.097
Question 2	L*	L+S	.057	.045	.032	**.925**	**.839**	**.935**	.019	.117	.032
Question 3	L*	S+S	.340	.283	.258	**.566**	**.619**	**.710**	.094	.099	.032
Question 4	L	L+L	.000	.022	.032	**.943**	**.915**	**.935**	.057	.063	.032
Question 5	L+S*	S+S	**.943**	**.816**	**.839**	.038	.045	.032	.019	.139	.129

*The options that were selected most per subsample are printed in bold face. The correct answers (i.e., the scenarios that were shown to be most effective by our calculations) are indicated with a *. There is no * in Question 4, since both situations contain an equal amount of expected bias.*

The three groups showed the same pattern in all five situations: participants preferred to use as much information as possible, i.e., they preferred the situation with the largest total sample size. For instance, the majority (57% of the students, 62% of the social scientists, 71% of the quantitative psychologists) preferred two small studies (total of 80 observations) over one large study (70 observations; Question 3). Second, most respondents believed that incorporating a small exact replication with a larger study in the evaluation (Question 2) would improve the accuracy of the estimate of the effect (93% of the students, 84% of the social scientists, 94% of the quantitative psychologists). So answers to questions 2 and 3 revealed two intuitions that are widely held among experts, social scientists, and students alike, namely, that (1) the larger the total sample size, the higher the accuracy, and (2) any replication, however small, improves accuracy. However logical these intuitions may appear at first sight, in this paper we show that both intuitions are false in the current publication system.

In this article we first explain the origin of these intuitions. Secondly, we show that replications are not science's Holy Grail, because of the 'replication paradox'; the publication of replications by itself does not decrease bias in effect size estimates. We show that this bias depends on sample size, population effect size, and publication bias. Finally, we discuss the implications for replications (and other studies that would be included in a meta-analysis of the effect under investigation) and consider possible solutions to problems associated with the use of multiple underpowered studies in the current publication system.

Why Do We Want More Observations and More Studies?

Our intuitions are grounded in what we learned in our first statistics courses, namely that: the larger the sample size, the more information, the greater the precision (i.e., the smaller the standard error), and the better the estimate. A replication study can also be viewed as increasing the original sample size. Hence, intuitively, both increasing the number of observations and incorporating a replication study increases the precision and the accuracy of the estimate of the population effect. This line of thought is reflected in the fact that multiple-study papers have increasingly become the norm in major psychology journals (Giner-Sorolla, 2012), although many of these involve conceptual replications rather than direct replications (Pashler & Harris, 2012; see also Makel et al., 2012).

Furthermore, there is also a large and growing literature on the merits of replication studies. For example, replications are said to be able to protect science from fraud and questionable research practices (Crocker & Cooper, 2011) and clarify ambiguous results (Simmons, Nelson, & Simonsohn, 2011). Replication is called “the gold standard for reliability” and “even if a small number of [independent replications] find the same result, then that result can be relied on” (Frank & Saxe, 2012). Finally, replications are supposed to uncover false positives that are the result of publication bias (Diekmann, 2011; Murayama, Pekrun, & Fiedler, 2013).

However, the above lines of reasoning do not take into account that publication bias may influence dissemination of both replication studies and original studies. We show how publication bias might limit the usefulness of replication studies and show why publication bias leads our intuitions and those of our colleagues (see Table 1) astray. We first present evidence of the omnipresence of publication bias in science, and show analytically how publication bias affects accuracy of the effect size estimate of a single study. Thereafter, we discuss the implications of our findings for the accuracy of effect size estimates in meta-analyses that include replications.

Publication Bias and How it Affects Effect Size Estimates

Presence of Publication Bias. Publication bias is the phenomenon that studies with results that are not statistically significant are less likely to be published (Greenwald, 1975). A way to search for publication bias is by looking for an overrepresentation of statistically significant or “positive” findings given the typical power of the studies (Ioannidis & Trikalinos, 2007). If there was no publication bias, and all effects were truly non-null (further called “true effects” or “existing effects”), then the proportion of positive findings in the literature would be approximately equal to the average power (the probability that you reject the null hypothesis when it is false). Although the recommended power for a study is at least .80 (e.g., Cohen, 1988), the median power has been estimated to average around .35 across studies in psychology (Bakker, van Dijk, & Wicherts, 2012)[2], the average power is .40-.47 across studies in behavioral ecology (Jennions & Moller, 2003)[3], and .21 across studies in neuroscience (Button et al., 2013)[4]. However, the rate of significant results is 95.1% in psychology and psychiatry, and 85% in neuroscience and behavior

(Fanelli, 2010). These numbers are incompatible with the average power across studies in the respective fields and represent strong evidence for publication bias in these fields.

An excess of significant findings has been established in many fields (Bakker et al., 2012; Button et al., 2013; Fanelli, 2012; Francis, 2014; Ioannidis, 2011; Kavvoura et al., 2008; Renkewitz, Fuchs, & Fiedler, 2011; Tsilidis, Papatheodorou, Evangelou, & Ioannidis, 2012). The rate of positive findings seems to be higher in the “softer” sciences, such as psychology, than in “harder” sciences, such as space sciences (Fanelli, 2010). There is evidence that the rate of positive findings has stayed approximately the same from the 1950’s (97.3% in psychology; Sterling, 1959) until the 1990s (95.6% in psychology and 85.4% in medical sciences; Sterling, Rosenbaum, & Weinkam, 1995), and that it even has increased since the 1990s (Fanelli, 2012).

Several studies have combined the results of tests of publication bias tests from multiple meta-analyses from various scientific fields and found evidence for publication bias in these fields. For instance, there is evidence for publication bias in about 10% of the meta-analyses in the field of genetic associations (Ioannidis, 2011), in roughly 15% of the meta-analyses in psychotherapy (Niemeyer, Musch, & Pietrowsky, 2012, 2013), in 20% to 40% of psychological meta-analyses (Ferguson & Brannick, 2012), in about 25%-50% of meta-analyses in the medical sciences (Sterne, Gavaghan, & Egger, 2000; Sutton, Duval, Tweedie, Abrams, & Jones, 2000), in 38%-50% of meta-analyses in ecology and evolution (Jennions & Moller, 2002), and in about 80% of meta-analyses in the field of communication sciences (Levine, Asada, & Carpenter, 2009). Although percentages of meta-analyses that are subject to publication bias do not seem to be impressively high, the power of publication bias tests was generally low in these meta-analyses. Hence, a failure to detect evidence for publication bias does not necessarily mean that there is no publication bias. A recent study established funnel plot asymmetry as a sign of publication bias in 82 meta-analyses (Fanelli & Ioannidis, 2013; see also Nuijten, Van Assen, Van Aert, & Wicherts, 2014).

Both the high prevalence of positive findings and the tests for publication bias in meta-analyses are not conclusive (but see Cooper, DeNeve, & Charlton, 1997; Franco, Malhotra, & Simonovits, 2014 for direct evidence of bias in psychology and the social sciences), but together they make a strong case for a presence of publication bias in much of the scientific literature. Therefore, it is important to investigate how studies are affected by publication bias.

The Effect of Publication Bias on an Estimate from a Single Study. We analytically derived the effect of publication bias on the effect size estimate in a published study with a two-independent samples design (see also Button et al., 2013; Gerber, Green, & Nickerson, 2001; Kraemer, Gardner, Brooks, & Yesavage, 1998). We used several scenarios differing in the degree of publication bias, the samples sizes, and the underlying effect size. Effect sizes were expressed in Cohen's *d*, or the standardized mean difference (i.e., $d = (\mu_1 - \mu_2)/\sigma$), with $\sigma = 1$). In each scenario we tested H_0: $d = 0$ against H_1: $d > 0$ using a *z* test. We also derived the effect of publication bias in the case where σ is unknown, using a *t*-test. Because the results of the two analyses are very similar, we only report those of the simpler *z* test. The equations and results for the *t*-test can be found at the Open Science Framework page https://osf.io/rumwi/.

We assumed that all significant results were published ($\alpha = .05$) and that there was one underlying effect. Two additional parameters were sample size *N,* and *pub*, representing the proportion of non-significant results published. We assumed that all non-significant *p*-values had the same probability of being published. Our assumptions on the probability of publication can also be interpreted differently, i.e., with *pub* as the probability of publication of a non-significant studies *relative* to the probability of publication of a significant study, where the latter probability can be smaller than 1. We were interested in the bias in the effect size estimate as a function of *d*, *pub*, and *N*. Figure 1 shows a variant of the typical depiction of power (used in most statistics textbooks) in which we display the effect of publication bias. Specifically, it shows the effect of *d* and *pub* on the published effect size estimate. In the figure "H0" and "H1" are the regions of accepting and rejecting the null hypothesis, respectively; 1-β represents power, α is the type I error, cv is the critical value of the *z* test, and *d* is the true population effect size. Without publication bias, available studies are drawn from the sampling distribution underlying *d* (H1). However, because of publication bias, non-significant results are less likely published, leading to an asymmetry of reported studies. Specifically, the dark gray area represents the proportion of studies with non-significant results that get published. The ratio of the lowered density (dark gray) to the regular density under H1 in the acceptance region equals *pub*, which equals .5 in Figure 1.

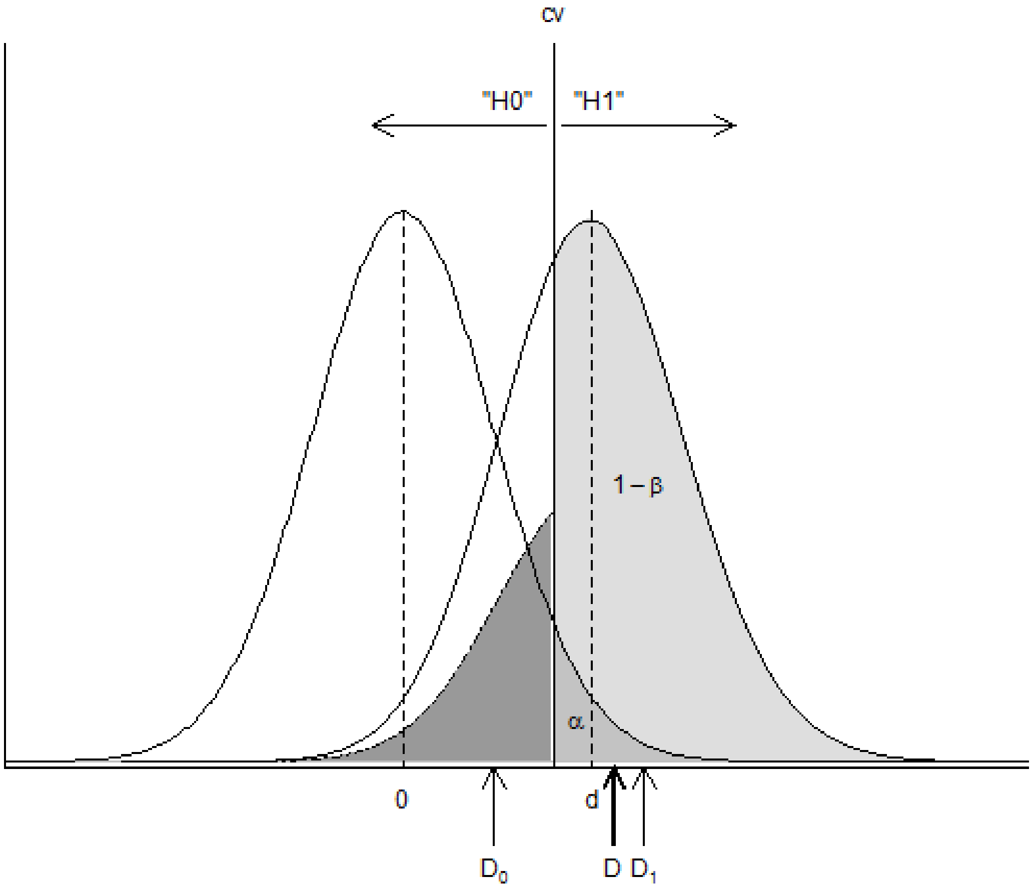

Figure 1. *Schematic representation of the effect of publication bias on the published effect size estimate. "H0" and "H1" are the regions of accepting and rejecting H0, respectively, 1-β represents power, α is the type I error, cv is the critical value of the test, and* d *is the true effect size.* D_0 *and* D_1 *are the expected effect sizes conditional on the acceptance or rejection of* H_0*, respectively, and D is the expected value of the published effect size.*

To establish the bias in the effect size estimate, we calculated the difference between the actual effect size *d*, and the expected value of the published effect size estimate, *D*. The value of *D* consists of two components. The first component is the expected value of the published effect size given that the effect size was significant, D_1, i.e., the expected value of the light-gray area. The second component is the expected value of the published effect size given that it was non-significant, D_0, or the expected value of the dark-gray area. The overall estimate *D* is a weighted average of D_1 and D_0, weighted by the light-gray and dark-gray areas, respectively. The higher the publication bias, the fewer non-significant findings are published, and the less weight D_0 will receive. In that case the weighted average will depend more on D_1, and D will overestimate d, as illustrated in Figure 1. If *pub* = 1 (no publication bias), the estimate D is equal to the true d, and if *pub* = 0 (maximal publication bias), the estimate D is equal to D_1, which overestimates d. Appendix 2 contains the exact equations.

In our analysis of the effect of publication bias on the accuracy of the effect size estimate in a published study we varied sample size (*N)* to be either 20 or 35 observations per group (40 or 70 observations in total, as in our questionnaire). These sample sizes were chosen to reflect typical sample sizes in psychology (Marszalek, Barber, Kohlhart, & Holmes, 2011; Wetzels et al., 2011). The population effect size, Cohen´s *d*, varied from zero to one. Finally, we chose values of *pub* equal to 0, .05, .25, .5, and 1. Values for *pub* of 0 and 1 reflect the two most extreme scenarios: total publication bias and no publication bias at all, respectively. The value .05 was based on an estimate of publication bias using the number of significant findings in the literature (see Appendix 3). We included the values .25 and .5 to reflect less severe publication bias. The dependent variable of our analysis is the bias in the effect size estimate, which is equal to the expected published effect size minus the true effect. The more bias in the effect size estimate, the less accurate the estimate. So in Figure 1, this amounts to the difference between *d* and *D*. Note that whereas this analysis renders the bias of the effect size estimate, the realized estimate will differ across studies and fields.

Figure 2 shows the effect of publication bias and population effect size on the bias in the effect size estimate in a single study with either 35 (left) or 20 observations per group (right). In both the large and the small study the same pattern appears. Both scenarios show that if the true effect size is sufficiently large, the bias approximates zero; the effect size estimate as it appears in the literature is equal to the true effect size. The nihil bias arises because for large enough effect sizes nearly all experiments are significant and therefore published. However, if the true effect size becomes smaller, more findings are non-significant and are not published. When that happens, bias or the overestimation of the effect generally increases.

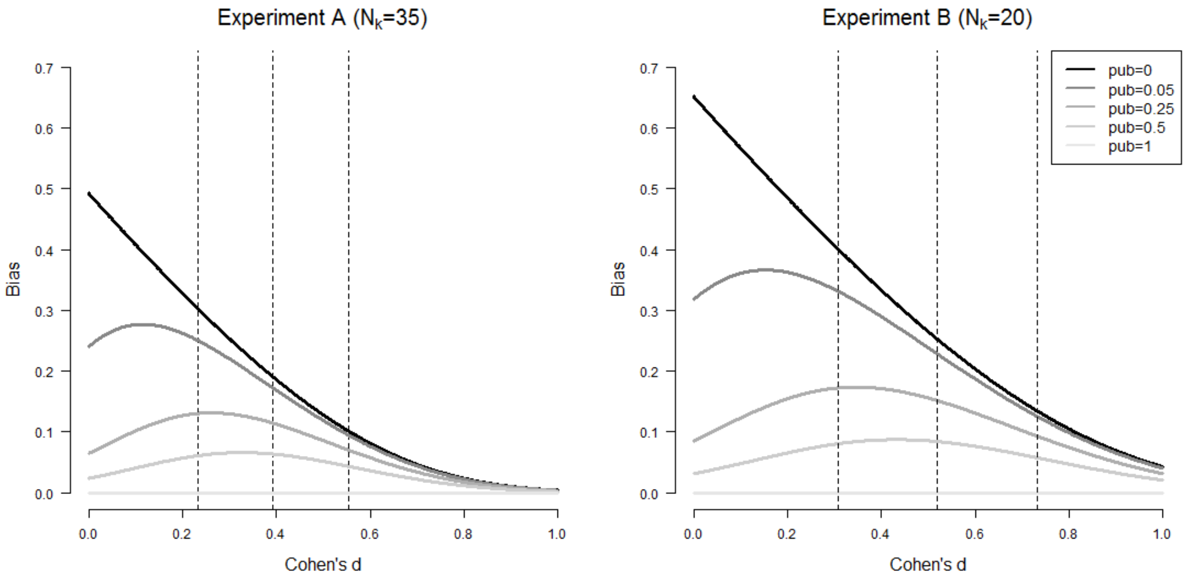

Figure 2. *The effect of publication bias and population effect size (Cohen's d) on the bias in the effect size estimate in a single study with either 35 or 20 observations per group. The bias in the effect size estimate is equal to the published effect minus the true effect. The vertical, dotted lines indicate Cohen's d at a power of .25, .50, and .75, respectively.*

Unsurprisingly, the magnitude of the bias depends on the severity of publication bias. If there is maximum publication bias (none of the non-significant results are published), the bias is the largest (black line in Figure 2). The bias decreases as more non-significant results are published. Without publication bias (results are published independent of their statistical significance), the bias in the effect size estimates disappears completely (lowest, light gray line in Figure 2). Formally, the (relative) bias compared to the situation where only significant results are published is a function of both *pub* and power (see Appendix 4 for the derivation of this equation):

$$Relative\ bias = \frac{1-pub}{1+pub\frac{\beta}{1-\beta}} \qquad (1)$$

It follows from (1) that bias already decreases dramatically for small values of *pub*, which is also apparent from the sharp drop in bias for *pub*=.05. For instance, consider a case in which *pub*=.05 and *d*=0. It follows that the obtained power is equal to α = .05. In this scenario we obtain a relative bias of (1-.05)/(1+.05*(.95/.05)) = .95/1.95 = .487, meaning that the bias is more than halved compared to the bias when *pub*=0. This is also apparent from Figure 2: in both the left and right panel it shows that at *d*=0 the bias in effect size estimate more than halves when *pub* increases from 0 to .05. Now consider a scenario where *pub* = .05 and power is .50 (middle vertical dotted line in Figure 2). Here we obtain a relative bias of (1-.05)/(1+.05*(.50/.50)) = .95/1.05 = .905, meaning that the bias is only slightly lower compared to the bias when *pub* = 0. It also follows from (1) that relative bias for a certain value of *pub* is only dependent on power. Hence both figures in Figure 2 have exactly the same shape. However, absolute bias decreases when sample size increases, hence bias is more severe in the small published study (right figure) than in the large published study (left figure). The difference in bias between the two studies is greatest when publication bias is maximal, and diminishes as publication bias decreases.

Surprisingly, Figure 2 shows that bias sometimes first *increases* when population effect size *d* increases. This happens whenever a small proportion of non-significant studies is published (*pub*=.05, .25, .5) and power is low. This somewhat counterintuitive result is due to two opposing forces. The first force is the decrease in bias for *pub* = 0 (upper black line); as *d* increases, the average D_1 of the light gray area in Figure 1 gets closer to *d*, thereby decreasing bias. The other force is relative bias; if *pub* > 0 and *d* increases, then power increases and relative bias (1) increases. Bias is the product of these two forces (see also Appendix 4). The bump in the figures for *pub* > 0 arises because the increase in relative bias overrules the decrease in bias for the significant studies whenever power is small. In other words, bias increases because the proportion of significant studies, which result in bias, increases more than their bias decreases as *d* increases. For larger values of power, bias decreases monotonically in *d* because then relative bias increases relatively less (see (1)) than bias for *pub* = 0 decreases.

The results of the analysis of the effect of publication bias and true effect size on the accuracy on effect size estimate when using a *t*-test (when σ is unknown) show that the shape of the figure based on the results of the *t*-test is identical to the shape of Figure 2.[5] The difference is that bias is slightly higher for the *t*-test than for the *z*-test, given the same publication bias and true effect size, and this difference decreases in sample size or degrees of freedom of the *t*-test.

An often-proposed solution to the problems of publication bias is to perform multiple studies within an article (see e.g., Murayama et al., 2013), or to add more replications (see e.g., Nosek, Spies, & Motyl, 2012). However, this advice does not take into account that such multiple studies may suffer from the same bias in effect size estimation because of publication bias

(Francis, 2012a). In the next paragraph we will therefore extend the known implications of publication bias on a single published study, to the implications of publication bias on scenarios with multiple published studies.

Implications of Publication Bias on the Accuracy of Multiple Published Studies

In this paragraph we show that replication studies are not necessarily a solution to the problem of overestimated effect size. In fact, we will show that replication can actually *add* bias to an effect size estimate under publication bias. We analytically derived the bias for three possible replication scenarios: two large studies, two small studies, and a large and a small study, and compared the bias in the effect size estimate with the bias in a single large study.

Let A be the original study, and B the replication. If we have two studies, the combined (weighted) effect size estimate *D* equals

$$\frac{N_A D_A + N_B D_B}{N_A + N_B}, \qquad (2)$$

where N_A and N_B represent the sample size, and D_A and D_B the estimated effect size of A and B, respectively. The results for the bias of estimated effect size based on both studies are shown in Figure 3.

The left panel of Figure 3 shows the bias after combining two large studies (one large study and a large replication). The responses to the questionnaire indicate that most researchers believe that two large studies yield a more accurate estimate of effect size than only one large study. However, the bias of two large studies is exactly the same as the bias in just one large study; because the replication contains the same amount of bias as the original study, the weighted average (2) of the two effect sizes will also contain the same amount of bias as the original study. Adding a replication to a single study will increase the *precision* or standard error of the estimate, but not its accuracy as long as there is publication bias.

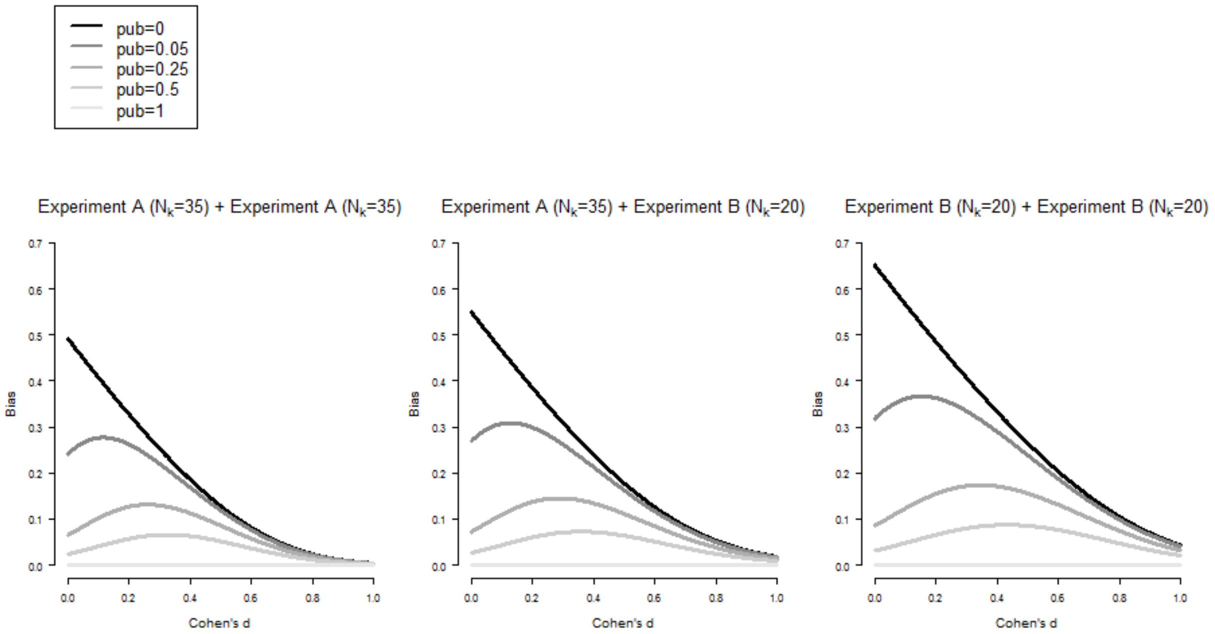

Figure 3. The effect of publication bias and population effect size (Cohen's d) on the bias in the effect size estimate in a replication scenario with either two large studies (left panel; identical to the bias in just one large study), one large and one small study (middle panel), or two small studies (right panel; identical to the bias in just one small study).

The middle panel of Figure 3 shows the bias in a large study combined with a small replication. According to the responses to the questionnaire, most researchers believe that a combination of one large and one small study yield a more accurate estimate than one large study. Again, this intuition is wrong when there is publication bias. Because a small study contains more bias than a large study, the weighted average (2) of the effect sizes in a large and a small study is more biased than the estimate in a single large study.

The right panel of Figure 3 shows the bias in a combination of two small studies. The responses to the questionnaire indicate that researchers believe a combination of two small published studies yields a more accurate estimate than one large published study. This intuition is not correct. Our analytical results show that the bias in the total effect size estimate does not change if effect size estimates of replication studies of the same size as the original study are synthesized with the effect size estimate of the original study. This means that the comparison between one large and two small studies is equivalent to a comparison between one large and one small study. Hence, the bias is larger in the combination of two small studies than in one large study, even though the sample size of the combination is larger than that of the large study.

In summary, in none of the three replication scenarios did the bias in the effect size estimate decrease by synthesizing the published replication with the large original published study. This means that both intuitions (1) the larger the total sample size, the higher the accuracy, and (2) any replication, however small, improves accuracy, are false when publication bias exists.

General Implications

Our examples and questionnaire refer to situations in which a published study is combined with a published exact replication. Our analysis shows that synthesizing a published original study with a published replication study generally does not decrease bias in the effect size estimate, yet may even increase bias if the replication study is smaller (in terms of sample size) than the original study. Our analysis has implications for more general situations such as combining effect size estimates of (i) an original study and a *larger* replication study (ii) published conceptual replication studies, (iii) conceptual replication studies within one single published article, (iv) many published studies on the same phenomenon, as in meta-analysis, and (v) for determining whether an effect exists or not.

In the light of recent calls for high-powered replication studies (see e.g., Brandt et al., 2014), we encounter more and more situations in which the replication study is actually larger than the original study. In those cases, the combined effect size estimate will have less bias than the effect size estimate of just the smaller, original study. Note, however, that in these cases incorporating the smaller original study in the estimation increases bias. Hence, evaluating only the large replication study would provide the most accurate effect size estimate (see also Kraemer et al., 1998).

The conclusion of our analysis holds for any situation in which two or more published effect sizes are combined to obtain an overall effect size (in a meta-analysis), when there is publication bias. This principle generally holds for all sample sizes, and any number of studies. The smaller the study, the larger the bias. So just like combining one small study with one larger study will increase bias in the effect size estimate, combining multiple smaller studies with multiple larger studies will also increase bias, as opposed to combining only large studies.

The same problem applies to situations in which conceptual (published) replications are combined to estimate one underlying (or average) effect size. If both the original study and its conceptual replication estimate the same population effect size and are subject to publication bias, both effect sizes will be inflated, and combining the two studies to obtain a new effect size will result in an overestimation of the population effect size, exactly in the same way as in our analysis. Similarly, the overestimation increases as the studies become smaller.

Multi-study papers are similarly affected by the paradox. Multiple studies within a single paper are also susceptible to publication bias (Francis, 2012b, 2012c, 2013; Francis, Tanzman, & Matthews, 2014), which means that an overall effect size based on the effects within one multi-study paper will be inflated as well. Our analysis generalizes straightforwardly to situations in which many published effect size estimates are combined, as in meta-analysis, which are also affected by publication bias (see e.g.,Fanelli & Ioannidis, 2013; Ferguson & Brannick, 2012; Nuijten et al., 2014). Here, too, overestimation gets worse whenever more small or underpowered published studies are included. What is even more problematic in meta-analysis is that *precision* of the effect size is increased (i.e., standard error of the estimate is decreased) by including more studies, thereby providing a false sense of security in the combined (biased) effect size estimate.

Publication bias also affects analyses used to establish whether an effect exists or not. It has been argued that replication may uncover false positives (e.g., Diekmann, 2011; Open Science Collaboration, 2012; Simmons et al., 2011), but this only holds if studies with non-significant results are accessible to researchers (see also Ferguson & Heene, 2012). Similarly, it has been argued that even though multi-study papers can inflate the effect size estimate, they can still decrease the rate of false positives (Murayama et al., 2013). The reasoning is that it is implausible that a research team generates, say, five false positive findings, since on average 5/.05 = 100 studies are needed to obtain five false positives. However, a problem in this argument is that the Type I error is typically much larger than .05, because of the use of so-called questionable research practices (QRP). For instance, Simmons et al. (2011) show that Type I error may even increase to .5 or higher after simultaneous use of some QRPs that are often used by researchers (John, Loewenstein, & Prelec, 2012; Simmons et al., 2011). Assuming a Type I error of about .5, five positive findings are no longer implausible, since only about ten studies need to be run. Both publication bias and QRP affect effect size estimates of smaller studies more than larger studies (Bakker et al., 2012; Fanelli & Ioannidis, 2013; Nuijten et al., 2014). This means that even if the goal is not to obtain an overall effect size, but to determine whether an effect exists, multiple underpowered published studies can still distort conclusions.

Does the problem of overestimation of population effect size also hold for unpublished research? We have to distinguish two different types of unpublished studies. First, there are unpublished studies, statistically significant or not, of which the results were subject to biases such as QRP. These biases result in overestimation of population effect size, even when a study’s outcome was not statistically significant (Bakker et al., 2012). This implies that incorporating these unpublished studies into meta-analyses may not decrease bias in effect size, particularly if their sample size is similar or smaller to those of published studies. Furthermore, this implication begs the question of the validity of publication bias tests that compare the effects of published and unpublished studies. These tests suggest there is no publication bias if the average effect sizes of published and unpublished studies are similar. Although this publication bias test addresses the effect of publication or not, a non-significant difference between the effects of published and unpublished studies does not imply that the published studies do not yield an overestimated effect size. Ferguson and Brannick (2012, p.126) even concluded that unpublished studies should not be included in meta-analyses, because searches for unpublished studies may be ineffective and unintentionally biased, and these studies may be inherently flawed. The second type of unpublished studies concerns studies that are not affected by biases such as QRP. Incorporating these studies into meta-analysis should generally decrease bias. However, these studies cannot or can hardly be distinguished from those unpublished studies affected by QRP as long as none of these studies are preregistered (see below). Because it is also unknown what proportion of unpublished studies is affected by QRP, it is impossible to tell to what extent unpublished studies yield overestimated effect sizes, both absolutely and relative to published studies.

Discussion

At the beginning of this article we presented results from a questionnaire that showed that psychology students, social scientists, and experts have the intuition that a published replication, independent of its sample size, improves accuracy of an estimated effect size. We also presented quotes from the published literature suggesting that replications are considered a tool to uncover false positives and to strengthen belief in true positives. We have shown that these intuitions do not hold in a publication system with substantial bias against non-significant results. The present system seems to be of this type, although some signs of improvement have recently emerged (e.g., Klein et al., 2014; Open Science Collaboration, 2012). We investigated the effect of replication on the bias in effect size estimate as a function of publication bias, sample size, and population effect size. We found that synthesizing a published original study with a published replication study can even add bias if the replication study's sample size is smaller than that of the original study, but only when there is publication bias. One implication of these findings is that replication studies are not necessarily the ultimate solution to false positives in the literature, as is sometimes implied, but should be evaluated with caution in the current publication system. Our results also hold more generally, i.e., for published conceptual replication studies, conceptual replication studies within one single published article, and many published studies on the same phenomenon, as in meta-analysis.

Our findings are based on the assumption that publication bias affects replication studies in the same way as it affects original studies. However, it is possible that this is not or no longer the case. For instance, publication bias might affect replications even more strongly than it affects original studies. Even though more and more psychologists have started to emphasize the advantages of replication studies, papers containing only one of more replications may still have a low probability of getting published (Giner-Sorolla, 2012; Makel, Plucker, & Hegarty, 2012; Neuliep & Crandall, 1990, 1993). Replications with non-significant results are easily dismissed with the argument that the replication might contain a confound that caused the null finding (Stroebe & Strack, 2014).

On the other hand, it is also possible that publication bias affects replications in the opposite way in some fields. That is, replications could have a *higher* chance of getting published if they contain non-significant results while a seminal study contains significant results, because this would be a controversial and thus an interesting finding. In that case, the next study would be controversial again if it were significant. What could follow is an alternation of effect sizes in opposite directions that eventually converge to – possibly – the true effect size. This is known as the Proteus phenomenon (Ioannidis & Trikalinos, 2005). If the Proteus phenomenon holds in practice, biased effect size estimates will cancel each other out over time and the overall effect size estimate will be close to unbiased (De Winter & Happee, 2013). Although the Proteus phenomenon may lead to unbiased effect size estimation, neglecting to publish studies with non-significant results is a very inefficient scientific enterprise with problems for statistical modeling of effect sizes (Van Assen, Van Aert, Nuijten, & Wicherts, 2014a, 2014b). Furthermore, even though there are occurrences of the Proteus phenomenon in some fields (Ioannidis, 2011), in psychology the vast majority of studies test if an effect is significantly different from zero, rather than if an effect is significantly different from a previously estimated effect (Fanelli, 2010, 2012; Van Assen, Van Aert, et al., 2014a).

Our analysis also assumes that there are no QRPs that affect the estimated effect size. Considering the seemingly widespread prevalence of QRPs (see e.g.,John et al., 2012), this might not be a realistic assumption. QRPs will likely also result in overestimation of effect sizes. Direct or close replication studies have generally less room for QRPs, since design, procedure, and measures are fixed by the original study. Hence less overestimation of effect size because of QRPs can be expected in direct replication studies. We must stress, however, that there exist only few studies of the effects of QRPs on effect size estimation, alone or in combination with publication bias (but see Bakker et al., 2012). Problematic is that QRPs are not well-defined and most likely have diverse effects on effect size estimation (cf. Lakens, in press).

There are several potential solutions to the problem of overestimation of effect sizes. The first solution is to only evaluate studies (and replications) with high precision or sample size (Stanley, Jarrell, & Doucouliagos, 2010) or, equivalently, high power. As our results showed, studies with high power will contain less bias in their effect size (see also Bakker et al., 2012; Button et al., 2013; Ioannidis, 2008; Kraemer et al., 1998). A related strategy is not only to evaluate, but also to conduct studies and replications with high power (Asendorpf et al., 2013; Brandt et al., 2014). Each of the studies with high power has little bias, and combining them will increase the precision of the final estimate. A complication with this solution, however, is that the power calculations cannot be based on the (previously) published effect size, because that published effect size is likely to be overestimated (see also Tversky & Kahneman, 1971). In order to perform an unbiased power calculation, the published effect size needs to be corrected for publication bias (Perugini, Galucci, & Constantini, 2014; Van Assen, Van Aert, & Wicherts, 2014; Vevea & Hedges, 1995).

A second solution is to eliminate publication bias altogether: without publication bias there is no bias in the effect size estimate. Many researchers have emphasized the importance of eliminating publication bias, and there are many proposals with plans of action. For instance, it has been proposed to split up the review process: reviewers should base their decision to accept or reject an article solely on the introduction and method section to ensure that the decision is independent of the outcome (Chambers, 2013; De Groot, 2014; Newcombe, 1987; Smulders, 2013; Walster & Cleary, 1970). A related method to eliminate publication bias is to evaluate submissions on their methodological rigor and not on their results. There are journals that evaluate all submissions according to these standards (see for instance PLoS ONE), journals with special sections for both "failed and successful" replication attempts (e.g., Journal of Experimental Social Psychology, Journal of Personality and Social Psychology, Psychological Science; Brandt et al., 2014), or websites like Psych File Drawer (http://psychfiledrawer.org) on which researchers can upload replication attempts. Furthermore, there have been large scale, preregistered replication attempts of different psychological experiments (Klein et al., 2014; Open Science Collaboration, 2012; see also Wagenmakers, Wetzels, Borsboom, Maas, & Kievit, 2012). However, even though these proposals and solutions show a high motivation to eliminate publication bias, finding and implementing the best strategy will take time.

What can we do with studies that are already published, and that most likely were subject to publication bias? Following upon others (e.g., Banks, Kepes, & Banks, 2012), we recommend publication bias analyses on past (as well as future) meta-analytic studies in an attempt to evaluate whether publication bias affected the estimated effect size in a field. Many different procedures exist that test for signs of publication bias (see e.g., Banks et al., 2012; Rothstein, Sutton, & Borenstein, 2005). A weakness of statistical procedures that test for publication bias, such as the rank correlation test (Begg & Mazumdar, 1994), Egger's test (Egger, Davey Smith, Schneider, & Minder, 1997), the trim and fill method (Duval & Tweedie, 2000a, 2000b), or Ioannidis and Trikalinos' test for an excess of significant findings (Ioannidis & Trikalinos, 2007; for an extensive discussion about this test and its usage see e.g., Ioannidis, 2013; Morey, 2013; Simonshon, 2013; Vandekerckhove, Guan, & Styrcula, 2013), is that their statistical power is usually low for meta-analyses with a typical number of studies. Consequently, when these procedures do not signal publication bias, publication bias may still be present and the meta-analysis' effect size estimate biased. On the other hand, these tests could also signal publication bias whenever there is none (a type I error). When this happens in a multi-study paper, the test would falsely imply that the author left out one or more studies, which may have unwarranted harmful consequences for the author.

Another option besides testing for publication bias is estimating an effect size that is robust against publication bias or one that is corrected for it. An often used procedure is the trim and fill method (Duval & Tweedie, 2000a, 2000b). However, the trim and fill method does not perform well with heterogeneous meta-analyses (Moreno et al., 2009; Terrin, Schmid, Lau, & Olkin, 2003) and its performance also depends strongly on assumptions about why studies are missing (Borenstein, Hedges, Higgins, & Rothstein, 2009). Another procedure that can be used to obtain unbiased effect sizes in the presence of publication bias is selection models (Copas, 2013; Hedges & Vevea, 1996, 2005; Vevea, Clements, & Hedges, 1993; Vevea & Hedges, 1995; Vevea & Woods, 2005). Selection models use the estimated or a priori probability that a study with a certain *p*-value is published, to estimate the influence of publication bias and to calculate an adjusted effect size. Selection models can deal with heterogeneous effect sizes (Hedges & Vevea, 2005), but may require many studies (e.g., 100 or more) to perform well (Field & Gillett, 2010). Furthermore, selection models are difficult to implement and depend on sophisticated choices and assumptions (Borenstein et al., 2009). A third procedure is to obtain an unbiased effect size by using only studies with statistically significant effects (Hedges, 1984; Simonsohn, Nelson, & Simmons, 2014; Van Assen, Van Aert, & Wicherts, 2014). Van Assen et al. (2014) show that their procedure, called *p*-uniform, provides unbiased effect size estimates, even with the relatively small number of eight studies in a meta-analysis, when the population effect size is homogenous. *p*-uniform also outperformed the standard fixed-effects meta-analysis, the trim and fill method, and the test of excess significance, when publication bias was present. Although we recognize the merits of all aforementioned procedures for testing and correcting for publication bias, they often lack power and/or require rather strong assumptions we believe these procedures do not provide the ultimate solution to problems resulting of publication bias.

Although we cannot establish the exact influence of publication bias on effect sizes estimates in published scientific articles, evidence suggests that publication bias affects many fields. To solve the problem of overestimated effect sizes, mere replication is not enough. Until there are ways to eliminate publication bias or correct for overestimation because of publication bias, researchers are wise to only incorporate and perform studies with high power, whether they are replications or not.

Footnotes

1. For more details about the sample and procedure, the original survey, the Dutch translation of the survey, and the full data set, see the Open Science Framework page https://osf.io/973mb/.
2. Estimated given a two independent samples comparison, assuming an effect size of *d* = .50 (based on estimates from meta-analyses) and a total sample size of 40, the median total sample size in psychology (Marszalek et al., 2011).
3. Based on 697 papers from 10 behavioral journals, assuming a medium effect size of *r* = .30. The authors report the estimated power for a small (r = .1), medium (r = .30), or large (r = .50) effect size. We report the power based on r = .30, because it is closest to the average estimated effect size in ecological or evolutionary studies of *r* = .18-.19 (based on 44 meta-analyses, Jennions & Moller, 2002). The average power we report here is therefore likely to be an optimistic estimate.
4. Based on data from 49 meta-analyses, using the estimated effect sizes in the meta-analyses as true effect sizes.
5. Equations and results for the *t* test can be found at the Open Science Framework page https://osf.io/rumwi/.

References

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108-119. doi: 10.1002/per.1919

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

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policy making. *Educational Evaluation and Policy Analysis, 34*, 259-277.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*(4), 1088-1101.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons, Ltd.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., . . . Van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217-224.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 1-12. doi: 10.1038/nrn3475

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*. doi: 10.1016/j.cortex.2012.12.016

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*: Routledge.

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

Copas, J. B. (2013). A likelihood-based sensitivity analysis for publication bias in meta-analysis. *Applied Statistics, 62*(1), 47-66.

Crocker, J., & Cooper, M. L. (2011). Addressing scientific fraud *Science, 334*, 1182. doi: 10.1126/science.1216775

De Groot, A. D. (2014). The meaning of “significance” for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta psychologica, 148*, 188-194.

De Winter, J., & Happee, R. (2013). Why selective publication of statistically significant results can be effective. *PLoS One, 8*, e66463. doi: 10.1371/journal.pone.0066463

Diekmann, A. (2011). Are most published research findings false? *Jahrbücher für Nationalökonomie und Statistik, 231*(5+6), 628-635.

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 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*, 455-463.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634.

Fanelli, D. (2010). "Positive" Results Increase Down the Hierarchy of the Sciences. *PLoS One, 5*(3), e10068. doi: 10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891-904. doi: 10.1007/s11192-011-0494-7

Fanelli, D., & Ioannidis, J. P. A. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences of the United States of America, 110*(37), 15031-15036.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*, 120-128. doi: 10.1037/a0024445

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: publication bias and psycholical science's aversion to the null. *Perspectives on Psychological Science, 7*(6), 555-561.

Field, A. P., & Gillett, R. (2010). How to do a meta‐analysis. *British Journal of Mathematical and Statistical Psychology, 63*(3), 665-694.

Francis, G. (2012a). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19*(6), 975-991. doi: 10.3758/s13423-012-0322-y

Francis, G. (2012b). The same old New Look: Publication bias in a study of wishful seeing. *i-Perception, 3*, 176-178. doi: 10.1068/i0519ic

Francis, G. (2012c). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review, 19*, 151-156. doi: 10.3758/s13423-012-0227-9

Francis, G. (2013). Publication bias in “Red, Rank, and Romance in Women Viewing Men” by Elliot et al. (2010). *Journal of Experimental Psychology: General, 142*, 292-296.

Francis, G. (2014). The Frequency of Excess Success for Articles in Psychological Science. *Psychonomic Bulletin & Review, 21*, 1180-1187.

Francis, G., Tanzman, J., & Matthews, W. J. (2014). Excess Success for Psychology Articles in the Journal Science. *PLoS One, 9*(12), e114255.

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

Frank, M. C., & Saxe, R. (2012). Teaching Replication. *Perspectives on Psychological Science, 7*(6), 600-604.

Gerber, A. S., Green, D. P., & Nickerson, D. (2001). Testing for Publication Bias in Political Science. *Political Analysis, 9*, 385-392.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7*, 562-571. doi: 10.1177/1745691612457576

Greenwald, A. G. (1975). Consequenceso f prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1-20.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*, 61-85.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*(4), 299-332.

Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 145-174). New York: Wiley.

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

Ioannidis, J. P. A. (2011). Excess Significance Bias in the Literature on Brain Volume Abnormalities. *Archives of General Psychiatry, 68*(8), 773-780.

Ioannidis, J. P. A. (2013). Clarifications on the application and interpretation of the test for excess significance and its extensions. *Journal of Mathematical Psychology, 57*(5), 184-187.

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology, 58*(6), 543-549. doi: 10.1016/j.jclinepi.2004.10.019

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4*(3), 245-253. doi: 10.1177/1740774507079441

Jennions, M. D., & Moller, A. P. (2002). Publication bias in ecology and evolution: an empirical assessment using the 'trim and fill' method. *Biological Reviews, 77*(2), 211-222. doi: 10.1017/s1464793101005875

Jennions, M. D., & Moller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology, 14*(3), 438-445.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological science, 23*, 524-532. doi: 10.1177/0956797611430953

Kavvoura, F. K., McQueen, M. B., Khoury, M. J., Tanzi, R. E., Bertram, L., & Ioannidis, J. P. A. (2008). Evaluation of the Potential Excess of Statistically Significant Findings in Published Genetic Association Studies: Application to Alzheimer's Disease. *American Journal of Epidemiology, 168*(8), 855-865. doi: 10.1093/aje/kwn206

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, J., Reginald B., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). *Investigating variation in replicability: A "Many Labs" Replication Project*. Retrieved from https://osf.io/wx7ck/

Kraemer, H. C., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3*, 23-31.

Lakens, D. (in press). What p-hacking really looks like: A comment on Masicampo & Lalande (2012). *Quarterly Journal of Experimental Psychology*.

Levine, T., Asada, K. J., & Carpenter, C. (2009). Sample Sizes and Effect Sizes are Negatively Correlated in Meta-Analyses: Evidence and Implications of a Publication Bias Against NonSignificant Findings. *Communication Monographs, 76*(3), 286-302. doi: 10.1080/03637750903074685

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research How Often Do They Really Occur? *Perspectives on Psychological Science, 7*(6), 537-542. doi: 10.1177/1745691612460688

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample Size in Psychological Research over the Past 30 Years. *Perceptual and Motor Skills, 112*(2), 331-348. doi: 10.2466/03.11.pms.112.2.331-348

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *Bmc Medical Research Methodology, 9*. doi: 2

10.1186/1471-2288-9-2

Morey, R. D. (2013). The consistency test does not–and cannot–deliver what is advertised: A comment on Francis (2013). *Journal of Mathematical Psychology, 57*(5), 180-183.

Murayama, K., Pekrun, R., & Fiedler, K. (2013). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*, 1088868313496330.

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5*(4), 85-90.

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality, 8*(6), 21-29.

Newcombe, R. G. (1987). Towards a reduction in publication bias. *British Medical Journal, 295*(6599), 656-659.

Niemeyer, H., Musch, J., & Pietrowsky, R. (2012). Publication bias in meta-analyses of the efficacy of psychotherapeutic interventions for schizophrenia. *Schizophrenia Research, 138*(2-3), 103-112. doi: 10.1016/j.schres.2012.03.023

Niemeyer, H., Musch, J., & Pietrowsky, R. (2013). Publication Bias in Meta-Analyses of the Efficacy of Psychotherapeutic Interventions for Depression. *Journal of Consulting and Clinical Psychology, 81*(1), 58-74. doi: 10.1037/a0031152

Nosek, B. A., Spies, J., & Motyl, M. (2012). Scientific Utopia: II - Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science, 7*, 615-631. doi: 10.1177/1745691612459058

Nuijten, M. B., Van Assen, M. A. L. M., Van Aert, R. C. M., & Wicherts, J. M. (2014). Standard analyses fail to show that US studies overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences, 111*(7), E712-E713.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657-660. doi: 10.1177/1745691612462588

Perugini, M., Galucci, M., & Constantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science, 9*(3), 319-332.

Renkewitz, F., Fuchs, H. M., & Fiedler, S. (2011). Is there evidence of publication biases in JDM research? *Judgment and Decision Making, 6*(8), 870-881.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis. Prevention, assessment, and adjustments*. New York: WIley.

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. doi: 10.1177/0956797611417632

Simonshon, U. (2013). It really just does not follow, comments on Francis (2013). *Journal of Mathematical Psychology, 57*(5), 174-176.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547.

Smulders, Y. M. (2013). A two-step manuscript submission process can reduce publication bias. *Journal of Clinical Epidemiology, 66*(9), 946-947. doi: 10.1016/j.jclinepi.2013.03.023

Stanley, T. D., Jarrell, S. B., & Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician, 64*(1), 70-77.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - Or vice versa. *Journal of the American Statistical Association, 54*, 30-34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited - The effect of the outcome of statistical tests on the decision to publish and vice-versa. *American Statistician, 49*(1), 108-112.

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology, 53*(11), 1119-1129. doi: 10.1016/S0895-4356(00)00242-0

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*(1), 59-71.

Sutton, A. J., Duval, S., Tweedie, R., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal, 320*(7249), 1574-1577.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*(13), 2113-2126. doi: 10.1002/sim.1461

Tsilidis, K. K., Papatheodorou, S. I., Evangelou, E., & Ioannidis, J. P. A. (2012). Evaluation of Excess Statistical Significance in Meta-analyses of 98 Biomarker Associations with Cancer Risk. *Journal of the National Cancer Institute, 104*(24), 1867-1878. doi: 10.1093/jnci/djs437

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

Van Assen, M. A. L. M., Van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014a). Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS One, 9*(1), e84896.

Van Assen, M. A. L. M., Van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014b). Why we need to publish all studies. from http://www.plosone.org/annotation/listThread.action?root=78405

Van Assen, M. A. L. M., Van Aert, R. C. M., & Wicherts, J. M. (2014). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*.

Vandekerckhove, J., Guan, M., & Styrcula, S. A. (2013). The consistency test may be too weak to be useful: Its systematic application would not improve effect size estimation in meta-analyses. *Journal of Mathematical Psychology, 57*(5), 170-173.

Vevea, J. L., Clements, N. C., & Hedges, L. V. (1993). Assessing the effects of selection bias on validity data for the General Aptitude-Test Battery. *Journal of Applied Psychology, 78*(6), 981-987.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*(3), 419-435.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10*(4), 428-443. doi: 10.1037/1082-989x.10.4.428

Wagenmakers, E. J., Wetzels, R., Borsboom, D., Maas, H. L. J. v. d., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632-638. doi: 10.1177/1745691612463078

Walster, G., & Cleary, T. (1970). A proposal for a new editorial policy in the social sciences. *American Statistician, 24*, 16-19.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*(3), 291-298.

Appendix 1: The survey including introduction text

The aim of this research is to examine how researchers value exact replications. More precisely, using five questions we assess your evaluation of **the effect of exact replication on the accuracy of the estimation of a population effect**. Accuracy is the closeness of the estimate to the population effect, and is inversely related to the bias of an estimate.

Introduction to questions: please read carefully

Imagine yourself being in the following situation. You want to estimate the effect of a treatment. To estimate this effect, you carry out a literature search. You only include **articles published in scientific journals** in your search. Additionally, you only include **exact replications** in your search. That is, the population, designs and procedures of the included studies are identical; the only difference between the exact replications may be their sample size. After your search you use the available empirical evidence to estimate the treatment effect in the population.

In the questions below you are asked to compare two situations. Your task in each question is to answer the question **‘Which situation yields the most accurate**

estimate of the effect of the treatment in the population?'. In both situations the same treatment effect is estimated. Hence, the question can also be formulated as **'Which situation would you prefer when your goal is to obtain an accurate estimate of the effect of the treatment in the population?'**.

A situation either involves one *published scientific article* (that is, no exact replications were found) or two *published scientific articles*. A published article is based on either **40** (**S**mall sample size) or **70** (**L**arge) observations. In the five questions below each situation is summarized by one or two letters. For instance, 'L' indicates that only one article was found with a sample size of 70. And 'L+S' indicates two studies were found that were exact replications of each other, one with 70 and the other with 40 observations.

Instruction for answering the questions

The table below contains both situations A and B of the questions (first columns) and the answers to the questions (last three columns). Answer the question by crossing *precisely one* of the three answering categories. For instance, consider Question 0 in the first row. Question 0 compares situation A and situation B, both with a small sample of 40 participants. The cross in the last column indicates that the respondent believes that both situations yield an equally accurate estimate of the effect of the treatment in the population.

Questions

Which situation (A or B) yields the most accurate estimate of the effect of the treatment in the population?

Question			Answer		
	Situation A	**Situation B**	**Situation A more accurate**	**Situation B more accurate**	**Situation A and B equally accurate**
Question 0	**S**	**S**			**X**
Question 1	**L**	**S**			
Question 2	**L**	**L+S**			
Question 3	**L**	**S+S**			
Question 4	**L**	**L+L**			
Question 5	**L+S**	**S+S**			

S = Small study with 40 observations; L = Large study with 70 observations

Thank you for your participation. Any questions or remarks about this research can be sent to Michèle Nuijten (m.b.nuijten@tilburguniversity.edu).

Appendix 2: Calculation of the Effect of Publication Bias and True Effect Size on the Accuracy on Effect Size Estimate When Using a *z*-test

The following equations show the influence of the proportion of non-significant results published (*pub*) on the accuracy of the effect size estimate in a single study, using a *z*-test comparing the means of two independent samples, with *σ* = 1 (see also Figure 1 for a schematic representation of these equations):

1) What is the critical value *cv* of the test?

$$cv = 1.645 \cdot \sqrt{2/_{N,}}$$

where *N* is the number of observations per group.

2) What is the z-value z_1 of the critical value under the alternative hypothesis?

$$z_1 = (cv - d) \cdot \sqrt{N/_2},$$

where *d* is the standardized true mean difference between the groups. The probability that $Z > z_1$ is the power of the test, 1-*β*.

3) What is the expected value D_1 of the mean difference, conditional on a rejection of H_0?

$$D_1 = \frac{f(z_1)}{(1-\beta) \cdot \sqrt{N/_2}} + d,$$

where $f(z_1)$ is the density of the standardized normal distribution at $z_{1.}$ The formula is based on the fact that the expected value of a truncated standardized normal distribution, truncated at probability *p*, equals $f(z_p)/(1-p)$.

4) What is the expected value D_0 of the mean difference, conditional on acceptation of H_0?

$$D_0 = d - \frac{f(z_1)}{\beta \cdot \sqrt{N/_2}}$$

Note that $\beta D_0 + (1\text{-}\beta)D_1 = d$, as it should.

5) What is the expected value *D* of the estimate of *d*?

$$D = \frac{pub\ \beta D_0 + (1 - \beta)D_1}{pub\ \beta + (1 - \beta)}$$

The derivations of our results using a *t*-test comparing the means of two independent samples are presented in an online Appendix at Open Science Framework: https://osf.io/rumwi/.

Appendix 3: Estimation of the Amount of Publication Bias in the Literature

We can make a rough estimate of the amount of publication bias in the literature based on the number of significant findings in the literature. We used the following equations (Van Assen, Van Aert, & Wicherts, 2014):

$$\mathrm{P}(\text{"}\mathrm{H}_1\text{"}|\mathrm{published}) = \frac{\mathrm{P}(\text{"}\mathrm{H}_1\text{"} \cap \mathrm{published})}{\mathrm{P}(\mathrm{published})} = \frac{\mathrm{P}(\text{"}\mathrm{H}_1\text{"} \cap \mathrm{published})}{\mathrm{P}(\text{"}\mathrm{H}_0\text{"} \cap \mathrm{published}) + \mathrm{P}(\text{"}\mathrm{H}_1\text{"} \cap \mathrm{published})}$$

$$= \frac{(1-\beta)\mathrm{P}(\mathrm{H}_1) + \alpha\,\mathrm{P}(\mathrm{H}_0)}{pub[\beta\,\mathrm{P}(\mathrm{H}_1) + (1-\alpha)\mathrm{P}(\mathrm{H}_0)] + (1-\beta)\mathrm{P}(\mathrm{H}_1) + \alpha\,\mathrm{P}(\mathrm{H}_0)},$$

where $\mathrm{P}(\text{"}\mathrm{H}_1\text{"})$ and $\mathrm{P}(\text{"}\mathrm{H}_0\text{"})$ are the proportion of significant and non-significant findings in the literature respectively, $\mathrm{P}(\mathrm{H}_1)$ and $\mathrm{P}(\mathrm{H}_0)$ are the proportion of effects that are truly non-null or null, respectively, α represents type I error, β represents type II error (and (1-β) represents power). Furthermore, *pub* < 1 represents the relative proportion of non-significant findings that are published, i.e. proportions of significant and insignificant findings that get published are assumed to be *q* and × *q*, respectively.

Following Ioannidis (2005), we assume that $\mathrm{P}(\mathrm{H}_1)$ is .50, which is perhaps an optimistic assumption, considering the exploratory nature of much psychological research. Furthermore, assume a power of .50 and α = .05. If we insert these values into the equation, and we assume that *pub* is .05, we get the following:

$$\mathrm{P}(\text{"}\mathrm{H}_1\text{"}|\mathrm{published}) = \frac{.5 \ast .5 + .05 \ast .5}{.05[.5 \ast .5 + (1-.05).5] + .5 \ast .5 + .05 \ast .5} = .88.$$

This result is in line with the research of Fanelli (2010) who found that between 84% and 91.5% of the papers in social and behavioral sciences report positive results. This would mean that the proportion of non-significant findings published lies around .05.

Of course this estimate of the amount of publication bias depends heavily on our assumptions. For instance, we could also consider a scenario in which α is not the nominal .05, but as high as .50. Simmons et al. (2011) indeed report that the actual α may increase from .05 to .5 when researchers employ several questionable research practices (QRP). When redoing our analysis with α = .5, with assuming these QRP will also boost power from .5 to .9, we obtain 88% reported positive results for *pub* = .32. To conclude, even when assuming scientists heavily use QRP, publication bias is estimated to be substantial.

Appendix 4: Calculation of Relative Bias in Effect Size Estimate

We can calculate the relative bias in effect size estimate compared to the situation where only significant results are published. Subtracting *d* from $D = \frac{pub\beta D_0 + (1-\beta)D_1}{pub\beta + (1-\beta)}$ yields the bias. Denote the bias for *pub* = 0, which equals $D_1 - d$, by *q*. Note that $D_0 - d = -\frac{1-\beta}{\beta}q$, since *d* is the weighted average of *D₀* and *D₁*, with type II error and power as weights, respectively.

Generally, for *pub* ≥ 0, bias $D - d$ can then be rewritten as

$$\frac{pub\beta D_0 + (1-\beta)D_1}{pub\beta + (1-\beta)} - \mathrm{d} = \frac{-pub(1-\beta)(D_1 - \mathrm{d}) + (1-\beta)(D_1 - \mathrm{d})}{pub\beta + (1-\beta)} = \mathrm{q}\frac{1-\mathrm{p}ub}{1+\mathrm{p}ub\frac{\beta}{1-\beta}},$$

where $\frac{1-\mathrm{p}ub}{1+\mathrm{p}ub\frac{\beta}{1-\beta}}$ denotes relative bias. This formula for relative bias also holds for the t-test.