The Psychology of Replication and Replication in Psychology

Gregory Francis

Perspectives on Psychological Science 2012 7: 585
DOI: 10.1177/1745691612459520

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(6) 585–594

DOI: 10.1177/1745691612459520
http://pps.sagepub.com
SAGE

The Psychology of Replication and Replication in Psychology

Gregory Francis
Department of Psychological Sciences, Purdue University

Abstract

Like other scientists, psychologists believe experimental replication to be the final arbiter for determining the validity of an empirical finding. Reports in psychology journals often attempt to prove the validity of a hypothesis or theory with multiple experiments that replicate a finding. Unfortunately, these efforts are sometimes misguided because in a field like experimental psychology, ever more successful replication does not necessarily ensure the validity of an empirical finding. When psychological experiments are analyzed with statistics, the rules of probability dictate that random samples should sometimes be selected that do not reject the null hypothesis, even if an effect is real. As a result, it is possible for a set of experiments to have too many successful replications. When there are too many successful replications for a given set of experiments, a skeptical scientist should be suspicious that null or negative findings have been suppressed, the experiments were run improperly, or the experiments were analyzed improperly. This article describes the implications of this observation and demonstrates how to test for too much successful replication by using a set of experiments from a recent research paper.

Keywords

aversion, effect size, memory, publication bias, power, replication, scientific method

It is probably safe to say that all scientists believe that empirical replication is a good thing. When someone doubts the validity of an experimental finding, a strong counterargument is to show that the finding successfully replicates in a new experiment. Throughout all of science and especially for fields that depend on statistical data analysis, leading researchers emphasize that experimental replication is the final arbiter in determining whether effects are true or false (Cohen, 1994; Fisher, 1935/1956; Roediger, 2012). There is much value to replication, but the belief that ever more successful replication verifies the validity of a finding is incorrect, because replication does not function in experimental psychology in the same way that it operates in some other scientific fields. Counterintuitive though it may seem, it is possible to have too much successful replication.

As shown below, the difficulties with replication in psychology are related to fundamental properties of hypothesis testing. Most experimental psychologists know that the process of hypothesis testing sometimes leads to Type I errors by rejecting a true null hypothesis. Indeed, this should happen around 5% of the time that the null is true, given conventional techniques. In a similar way, experiments sometimes make Type II errors by not rejecting a false null hypothesis. The probability of a Type II error depends on the size of the effect, the design of the experiment, and the experiment sample size(s). Psychologists seem to forget about Type II errors, perhaps because they are difficult to estimate. The existence of Type II error (and its complement, power) implies that even if an effect is real, some experiments should fail to reject the null hypothesis. Another way to describe this property is that failures to replicate should occur with some probability, even when the effect is true.

More precisely, even experiments measuring a true nonzero effect should successfully replicate the existence of that effect only at a rate that is consistent with the power of the experiments. Ioannidis and Trikalinos (2007) showed how to use this fundamental characteristic of hypothesis testing to identify an excess of significant results, which can be interpreted as the presence of publication bias. When there are significantly more reports of experiments that reject the null hypothesis than is consistent with a power analysis of those experiments, there is evidence either that some null or negative findings were suppressed or that the reported experiments were run improperly.

Corresponding Author:
Gregory Francis, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47906
E-mail: gfrancis@purdue.edu

The Ioannidis and Trikalinos publication bias test has started to be applied to investigations in psychology. Renkewitz, Fuchs, and Fiedler (2011) used it to identify publication bias in two meta-analyses related to judgment and decision making. Francis (2012a, 2012b, 2012c, 2012d, in press) used the test to demonstrate that publication bias appeared to have contaminated some sets of findings in experimental psychology, including the precognition studies of Bem (2011). The presence of publication bias does not necessarily mean a reported effect is false, but it suggests that the presented evidence is unconvincing. In a sense, an experiment set contaminated with publication bias is anecdotal and therefore nonscientific. This analysis suggests that there is no reason to believe Bem's claim that people can sense the future and use that information to guide present actions. Schimmack (in press) used a similar analysis to reach the same conclusion.

To explain these ideas, the following section describes an analysis of a set of experiments reported by Galak and Meyvis (2011), who found that people remembered an unpleasant event as being more aversive if they believed they would soon experience the event again. In this one report, the main finding was replicated multiple times. As the analysis below shows, given the properties of the experiments in Galak and Meyvis (2011), the reported number of successful replications is suspicious.

Before going through the analysis, it is important to emphasize that evidence of publication bias does not necessarily mean that a reported phenomenon is false. Rather, the presence of publication bias means that a set of reported experiments does not appropriately investigate the validity of the phenomenon. An effect may be real or not, but a set of experiments with publication bias does not clarify the situation; instead, additional unbiased experiments are required to determine whether such an effect is real.

Also the presence of publication bias does not mean that it is appropriate to conclude that authors deliberately set out to mislead readers. As argued below, publication bias could occur in a set of experiments because authors closely follow the standards of the field. Indeed, the experiments reported in Galak and Meyvis (2011) appear to meet (and often exceed) the standards of experimental psychology. The quality of the report (by current standards) is related to the strength of evidence for publication bias.

Analysis of Galak and Meyvis's (2011) Experiments

Galak and Meyvis (2011) showed that anticipation of a repeat exposure to an unpleasant situation (e.g., a boring task) leads to higher aversive ratings for the memory of the previous exposure. Table 1 shows the main statistical properties of eight experiments that consistently showed this effect. Each experiment tested differences between aversive ratings of a previous unpleasant situation for a group of participants who believed they would experience more of the unpleasant situation and for a group of participants who believed they were done with the unpleasant situation. The main finding was that the aversive ratings were higher for people in the more exposure group compared with the done group. There were additional studies that measured a conceptually similar effect in a nonlaboratory situation (e.g., women's ratings of menstruation and runners' ratings of climbing a hill), but they are not included in the analysis because they involved comparisons that cannot be directly matched with the findings in Table 1.

Table 1. Statistical Properties of the Galak and Meyvis (2011) Experiments.

Description	n_1	n_2	Effect size (g)	Power from pooled ES
Study 1a	15	15	0.745	0.257
Study 1b	15	15	0.832	
Study 2	22	22	0.688	0.452
Study 3	28	28	0.657	0.556
Study 4	52	51	0.377	0.815
Study 5	44	43	0.652	0.747
Study 6	41	41	0.482	0.725
Study 7	25	26	0.628	0.511

Note: Effect sizes (ESs) were computed from the reported F values. The power for Study 1 is for both findings rejecting the null hypothesis.

Effect sizes

Each statistical analysis was based on an analysis of covariance (ANCOVA) that factored out a participant's initial aversive rating for the unpleasant situation. The authors noted that there was never a statistical difference between the two groups for these initial ratings, so the ANCOVA factorization probably makes little difference in the data analysis, and a normal analysis of variance would have likely produced similar results. The overall effect of the difference between the two groups can be estimated with a meta-analytic technique that pools the effect sizes across all of the experiments. The first step is to compute a standardized effect size for each experiment, Hedges g. The formula is

$$g = J\sqrt{\frac{F\left(n_1+n_2\right)\left(1-R^2\right)}{n_1 n_2}} \quad (1)$$

where F is the reported statistic, R is the covariate outcome correlation, and $J = 1-3/[4(n_1 + n_2 - 2) - 1]$ is a correction factor for small sample sizes (Hedges & Olkin, 1985). This effect size describes the difference between the two groups in terms of the standard deviation of the data. Bigger effect sizes indicate stronger effects on the judgments of remembered aversion as a result of anticipated (or not) additional exposure to the unpleasant context. Galak and Meyvis (2011) did not report R, so for the present analysis it was set to zero, which maximizes g (and is generally consistent with their observation that there was no

difference in initial aversion ratings between the two groups). This choice for R means that the effect size values in the fourth column of Table 1 tend to overestimate the true effect sizes.

In a meta-analysis, the experiment-wise effect sizes are combined (by weighting with the inverse variance of the effect size) to create a pooled effect size (Hedges & Olkin, 1985). Study 1 is a special case because one set of participants generated the two reported effect sizes. To deal with dependency between these effect sizes, they were averaged, and then the averaged effect size was entered into the meta-analytic pooling calculation. The pooled effect size across all of the experiments, $g^* = 0.568$, is the best estimate of the standardized effect size, and it can be used to compute the power of each experiment.

Power

Power is the probability that an experiment will reject the null hypothesis for a given effect size. Experiments with larger effects or larger sample sizes have more power.

The last column of Table 1 shows the power of each experiment using the pooled effect size. For simple cases, power can be estimated with relatively straightforward calculations that use the effect size and sample sizes (Champely, 2009; R Development Core Team, 2011). A special case is that the power value given for Study 1 in Table 1 is the probability of rejecting the null hypothesis for both hypothesis tests. This value was estimated with a simulation of 100,000 experiments. In each simulation, the sampling populations from the two measurements were correlated with $r = .9$ (a larger correlation would give a larger power, but it would never go above .324, which is the power of one experiment by itself). The simulations also computed the probability of none of the tests rejecting the null hypothesis (.608) and each test alone (.068 and .066).

The power values for Studies 3, 4, 5, and 6 were also calculated with simulations because the standard deviation was estimated using Fisher's least significant difference (LSD) test, so the degrees of freedom in the tests are larger than what is indicated by the sample sizes in Table 1. The LSD test makes each experiment slightly more powerful than it would be normally.

Publication bias

The sum of the power values (4.06) is the expected number of times eight experiments like these should reject the null hypothesis. With a criterion of $p = .05$, seven of the eight experiments actually rejected the null hypothesis. The probability that seven or more experiments like these would reject the null hypothesis is computed with an exact test that considers all nine possible combinations of experiments with seven or eight rejections. The probability of each combination was found by multiplying the power and Type II error values as appropriate for each experiment. These probabilities were then summed to give the overall probability, .079, of observing seven (or more) experiments that reject the null hypothesis. This probability is below the .1 criterion that is typically used to indicate publication bias (Begg & Mazumdar, 1994; Ioannidis & Trikalinos, 2007; Sterne, Gavaghan, & Egger, 2000). Thus, the number of successful replications of the effect is surprising, given the size of the effect and the experiments that measured it.

It might be tempting to argue that the probability of the experimental findings is not very much below the .1 criterion (which is admittedly somewhat arbitrary).[1] However, because the effect sizes were computed with $R = 0$ in Equation 1, they were likely overestimated. In addition, as shown below, a common side effect of publication bias is an exaggeration of reported effect sizes. The overestimated effect sizes lead to overestimated power values, so the bias problem is likely worse than what the test indicates.

Given the misunderstandings about replication in psychology, there is strong pressure for authors to present only statistically significant findings. This pressure can lead to strange choices in how authors present their results, and these choices often exacerbate the conclusion of publication bias. For example, the above analysis deviated a bit from the interpretation of the findings made by Galak and Meyvis (2011). Power depends on the criterion value that is used to determine statistical significance. By convention, this criterion is set to be .05, which defines the probability of a Type I error (rejecting the null hypothesis when it was really true). Galak and Meyvis (2011) usually followed the convention, but for Study 4 they computed $p = .056$ and concluded that they had replicated their finding. Strictly following convention, such a finding would be treated as a failure to replicate, and this is how it was interpreted in the above analysis.

However, the criterion p value is somewhat arbitrary, and perhaps Galak and Meyvis (2011) had valid reasons to use a nonstandard criterion (say, $p = .06$) as a basis for rejecting the null hypothesis. If we imagine the criterion to be this nonstandard value, then the power for Study 4 is a bit larger, 0.837. However, under this interpretation, all eight experiments rejected the null hypothesis, and the probability of this happening is the product of the power values, which is 0.015. Thus, by adjusting the significance criterion to make it so that all of the experiments rejected the null hypothesis, Galak and Meyvis (2011) made the experiments very unbelievable as a set.

In an important sense, Galak and Meyvis (2011) are correct that Study 4 provides evidence for their effect. Their mistake was to suppose that statistical significance was needed to demonstrate such evidence. If there were no publication bias, a meta-analysis would properly use the nonsignificant finding in Study 4 as useful information about the pooled effect size.

The problems with the biased reported findings are not alleviated by the additional conceptual replications in the report. In addition to the findings reported in Table 1, Galak and Meyvis (2011) had at least three successful conceptual replications of the same effect. Because they measured somewhat different phenomena, these conceptual replications cannot be pooled

with the effect sizes in Table 1, but they still have power values of less than one. It is difficult to accurately measure power from a single experiment (Yuan & Maxwell, 2005), but if one imagines that the three additional experiments each had power values of 0.85, which is larger than any of the other experiments, then the probability of having 10 out of 11 experiments reject the null hypothesis is .053. The probability of all 11 experiments rejecting the null hypothesis (using the nonstandard criterion in Study 4) is .009.

Publication bias and theory predictions

A similar analysis can be used to consider the strength of the evidence in support of Galak and Meyvis's (2011) theory. Their conclusions emphasized that the theory was able to distinguish experiments that should, and should not, show the effect. They noted that in addition to the repeated success for predicting experiments to reject the null hypothesis, the theory also successfully predicted nine additional findings where the aversion ratings were not affected by an anticipated return to an unpleasant context. It may seem that the ability of the theory to predict hits and misses so accurately is strong validation of the theory, but this interpretation is misguided because the experiments lack sufficient power to support such claims. Just as there can be too much successful replication, excessive validation of a theory can also indicate a problem.

Suppose that the theory is correct and that the situations where it predicts no effect really have an effect size of zero. As a result of random sampling, each experiment has a .05 probability of rejecting the null hypothesis. Thus, the probability that nine such experiments do not reject the null hypothesis is $(1 - .05)^9 = .63$. The previous analysis established that the probability of the theory correctly predicting all but one of the rejections of the null hypothesis for the eight experiments described in Table 1 is likely no more than .079. The probability of the theory being so accurate that it correctly predicts all but one of the reported rejections and all of the reported nonrejections of the null hypothesis is no more than the product of these two probabilities, which is .050. (It is no more than .009 if one uses the nonstandard rejection criterion for Study 4.) Another way to describe the conclusion of this analysis is that it is not believable that a theory should be so accurate when the experiments should show so much uncertainty.

In one sense, it is not the theory that is to blame for this conclusion; the fault is with the experiments. Given the effect size estimated by the experiments, the studies in Galak and Meyvis (2011) are underpowered. Two of the experiments had power values less than 0.5, and two other experiments had power values below 0.6, but every one of these experiments rejected the null hypothesis. Only three experiments had power values greater than 0.7. One cannot draw strong inferences from any specific pattern, but it is curious that the experiment with the highest power (Study 4) was the one experiment that failed to reject the null hypothesis with a .05 criterion. Given these relatively low power values, it will not be possible to use the outcome of these experiments to draw firm conclusions about any theory. Roberts and Pashler (2000) made somewhat similar points about conclusions that can be drawn given the level of uncertainty in a model and in data.

However, the nature and use of the theory proposed by Galak and Meyvis (2011) reflects the general confusion psychologists appear to have about replication. As is common in psychology, they described their theory in a nonquantitative way. It consists of some ideas that predict relationships between variables, and all of these relationships are described at a verbal level. As is also common in psychology, they used the theory to predict null hypothesis rejections and nonrejections in a way that does not consider the design or power of the experiments. Although common, this approach ignores the fundamental properties of hypothesis testing. No theory (verbal or quantitative) should directly predict the outcome of a hypothesis test because a predicted experimental outcome is not well defined without first specifying the properties (e.g., sample sizes, effect size, power) of the test.

The Importance of Effect Sizes

Rather than focusing on whether an experiment rejects the null hypothesis, psychologists should use experiments to characterize the magnitude of effect sizes. Likewise, theories should not bother with verbal descriptions of how variables are related to each other but instead should quantitatively predict effect sizes and relationships between effect sizes. With a theory of effect sizes, psychologists can easily predict the power of experiments and then design experiments that properly test a theory's predictions. Psychologists need to recognize that effect size estimation (magnitude and precision) is the key property of their experiments, and this is true whether one uses traditional methods or Bayesian approaches (Cumming, 2012; Kruschke, 2010; Rouder, Speckman, Sun, Morey, & Iverson, 2009). There is a role for hypothesis testing in psychology, but it is not necessary for many experimental studies.

Given that almost every empirical study in psychology currently uses hypothesis testing, it may seem bizarre to claim that effect sizes are more important than the outcome of hypothesis tests; but the case can be argued in at least two ways. First, as noted above, effect sizes play a central role in predicting the outcome of hypothesis testing. Surely, the theory of Galak and Meyvis (2011) does not predict that every experiment testing the theory will reject the null (e.g., even with sample sizes of, say, $n = 3$). The authors probably meant that if the theory were true, then an experiment with a large enough sample would reject the null hypothesis. But the definition of "large enough sample" is determined by the magnitude of the effect size, so even if you believe that the goal of experiments is to indicate whether an effect exists (the outcome of a hypothesis test), you still have to focus on effect sizes in order to predict the outcome of experiments. If you do not have any estimate of an effect size (through prior work or theory), then before predicting the outcome of a hypothesis

test, you need to gather data to estimate the effect size. Note, even with a given effect size, the best a researcher can do is estimate the probability of rejecting the null; there is always some uncertainty in experimental outcomes.

Second, the importance of effect size measurement is reflected in the fact that almost every quantitative model in psychology describes effect measures, such as reaction time, proportion correct, contrast threshold, or some function of such effects. For example, Dutilh et al. (2012) compared model predicted and empirical data for posterror slowing (differences in reaction time for posterror and postcorrect trials). The reported output of the model is an effect (difference in reaction times), not a statement about statistical significance. Indeed, it is difficult to find any situation where a quantitative model explicitly focuses on the outcome of a hypothesis test rather than the magnitude (and precision) of an effect. Empirical studies should focus on effect sizes because they are important for developing quantitative theories that summarize empirical data and provide a framework for understanding psychological mechanisms. The outcome of a hypothesis test is mostly irrelevant for such modeling efforts.

How Does Publication Bias Happen?

By the current standards of science and scientific reporting, Galak and Meyvis (2011) is an exemplary article. The findings are grounded in theory, connected to other phenomena, successfully replicated many times, and explained with an intuitively plausible theory that accurately predicts both the presence and absence of an effect. Nevertheless, the analysis above demonstrates that it is precisely because the article meets (and exceeds) the current standards of psychological science that there is strong evidence of publication bias. The troubling implication is that the standards of science in experimental psychology are flawed. It seems likely that similar problems apply to many other findings in experimental psychology (and in other fields that depend on statistics), although the above analysis may not be applicable if a phenomenon has not had many replication attempts.

Given the low probability that the experiments in Galak and Meyvis (2011) would produce the number of reported null hypothesis rejections, it is valuable to consider how this could have happened. There are four broad possibilities:

1. *Chance*: Every decision-making process has a risk of making a mistake in an uncertain environment. It is possible that a set of experiments is not actually biased but just happens to produce unusual samples that appear to be biased. Regardless of this risk, the proper scientific interpretation of a set of apparently biased findings is to doubt the validity of the experiments. To do otherwise is to reject the tenets of hypothesis testing.
2. *File-drawer problem*: There may have been additional experiments that did not reject the null hypothesis and were not described in the published report. This could be because the authors chose to not describe such experiments or because reviewers or the editor asked them to remove the null findings before the manuscript could be published.
3. *Inflated frequency of rejecting the null hypothesis*: The experiments may have been run improperly in a way that inflated the rate of rejection of the null hypothesis. Simmons, Nelson, and Simonsohn (2011) described several tricks that can inflate the rejection rate of an experiment (regardless of whether the null hypothesis is true or false). Rejection rate inflation can also occur because of improper analytic techniques (McCullough & McWilliams, 2010, 2011).
4. *Underestimation of the true effect size*: The experiments may have been run improperly in a way that underestimated the effect size. With an underestimated effect size, the above analysis will underestimate the power of the experiments. Ioannidis (2008) notes that effect size underestimation can happen for some sequential sampling situations when the true effect size is large because an experiment can reject the null hypothesis with an underestimated value of the effect size. Simulation examples of this situation will be described in the discussion of Figure 1B. Even for such cases, the underestimation tends to be small, so this explanation is unlikely to account for the low believability of the findings in Galak and Meyvis (2011).

There does not appear to be a method for identifying which of these broad explanations (and it may be more than one) contribute to the appearance of publication bias in a set of experiments.

Properties of Publication Bias

This section uses multiple simulated two-sample *t* tests to demonstrate general properties and characteristics of publication bias. Each simulated experiment took a random sample of 30 data points from a standard normal distribution (mean of zero and standard deviation of one) as a control group and a random sample of 30 data points from a normal distribution with a mean of *g* and a standard deviation of one as an experimental group. Thus, the true effect size for the experiment was *g*. A two-sample, two-tailed *t* test with $\alpha = .05$ was used to determine statistical significance. A set of 10 such experiments with data taken from populations with the same true effect size, *g*, was used to compute the meta-analytic pooled effect size, g^*. The light gray dots in Figure 1A show the pooled effect size plotted against the true effect size, for true effect size values between zero and 1.5.

The simulation was repeated 30 times for each true effect size. Because meta-analysis works, it should not be surprising that the light gray dots cluster around the black diagonal line,

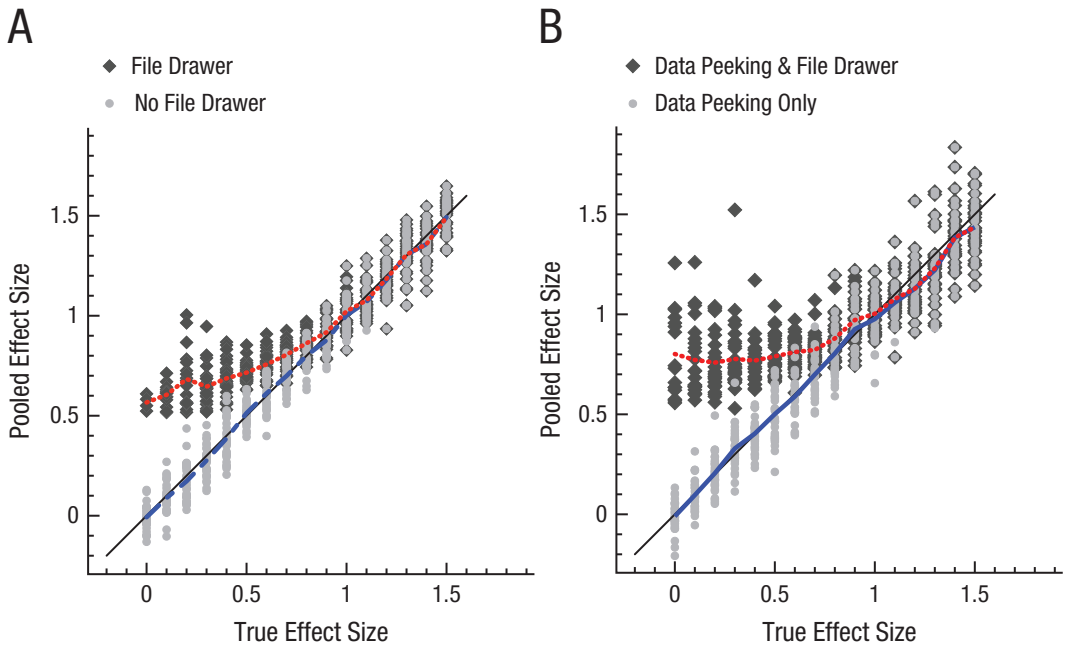

Fig. 1. Pooled effect sizes computed from simulated experiments under various types of publication bias. In Panel A, each light gray circle corresponds to the meta-analytic pooled effect size from a set of properly run 10 experiments without any publication bias. Each dark gray diamond is the pooled effect size of only those experiments that reject the null hypothesis. In Panel B, the same kind of analysis is performed, except all of the experiments use a data peeking strategy. In terms of estimating the effect size, the findings are very similar.

which corresponds to the true effect size. The dashed line plots the mean of g^* across the 30 sets of experiments for each true effect size. The meta-analysis works properly because the entire set of findings has been fully reported, regardless of whether an experiment rejected the null hypothesis.

File-drawer bias

The dark gray diamonds in Figure 1A describe the values of the same type of meta-analysis of effect size, but the set of experiments was subjected to a file-drawer bias that published only those experiments that rejected the null hypothesis in a positive direction (i.e., the experimental group mean is bigger than the control group mean). Thus, only those experiments that rejected the null hypothesis (in a positive direction) were part of the meta-analysis. Because only experiments with relatively large experiment-wise effect sizes can reject the null hypothesis, the file-drawer bias caused the pooled effect size to grossly overestimate the true effect size. Note, even when the true effect size was zero, the pooled effect size of the reported experiments was usually above 0.5. The dotted line reports the mean pooled effect size across the simulations for each true effect size. That the dotted line converges on the dashed line for large true effect sizes simply reflects the property that almost all of the experiments rejected the null hypothesis when the true effect size was large enough. When some experiments do not reject the null hypothesis, a file-drawer bias mischaracterizes the magnitude of an effect.

It may seem unlikely that anyone would use a file-drawer bias for a small true effect size because so few experiments would reject the null hypothesis. When the true effect size is zero (the null hypothesis is really true), usually none of the 10 simulated experiments reject the null hypothesis. Only around 2.5% of the time does an experiment reject the null hypothesis in the desired direction (just by random sampling). Few researchers would deliberately suppress nine (or more) null findings in order to publish one experiment that rejected the null hypothesis, if only because it is a very inefficient way of producing significant experimental findings. Such efforts would be exhausting and properly characterized as fraud. However, there are ways to (improperly) increase the rate of rejecting the null hypothesis.

Data peeking

Consider a sequential sampling approach that is sometimes called data peeking (or optional stopping); here the experimenter runs a hypothesis test while gathering data and stops the experiment when the null hypothesis is rejected or when the sample size reaches an upper limit (Berger & Berry, 1988; Strube, 2006). New simulated experiments were run with a data peeking strategy. Each experiment started with a sample size of 10 for both the control and experimental groups. If the null hypothesis was rejected with a t test, the experiment was stopped. If the null hypothesis was not rejected, one additional data point was selected for each group and another t test was

run. This process continued until the experiment stopped or the sample size reached a maximum of 30 data points. The light gray dots in Figure 1B show the pooled effect size, for a set of 10 simulated experiments, plotted against the true effect size. The pooled effect size is not greatly affected by data peeking. (The slight underestimation of the effect size for large true effect size values is due to stopping the experiment early when rejecting the null hypothesis. Such stopping can occur for effect sizes that are smaller than the true value, which biases the pooled effect size.) However, the introduction of an additional file-drawer bias to the experiment set leads to massive overestimation of the pooled effect size for a set of experiments, as indicated by the dark gray diamonds in Figure 1B.

In many respects, the findings in Figures 1A and B are very similar. However, what has fundamentally changed between the data peeking results in Figure 1B and the non–data peeking results in Figure 1A is that the data peeking approach increased the frequency of rejecting the null hypothesis. Figure 2 plots the mean number of experiments (out of 10) that rejected the null hypothesis for the experiment sets that produced Figure 1. With data peeking, the mean number of experiments that rejected the null hypothesis was larger than without data peeking. Data peeking by itself introduces a publication bias by giving a false sense of how often an effect will reject the null hypothesis. An equally important property of data peeking is that it makes it easier to implement a file-drawer bias. Consider a true effect size of 0.3 in Figure 2. Without data peeking, a set of 10 experiments with a sample size of 30 data points in each group will reject the null hypothesis only about 1.7 times. If the experiments were run with data peeking (starting from 10 and going to 30 data points in each group), there will be almost four rejections of the null hypothesis. Even when the null hypothesis is true, data peeking increases the rejection rate (now Type I error) from around 0.02 up to almost one experiment in 10.

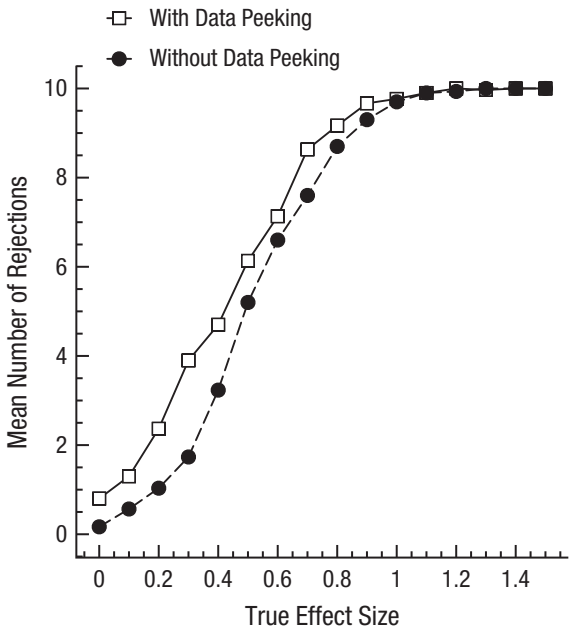

Fig. 2. For the simulated experiment sets reported in Figure 1, using data peeking leads to more experiments that reject the null hypothesis.

The increase in the rejection rate introduced by data peeking may not seem like much, but Simmons et al. (2011) described several other tricks that can also inflate the rejection rate, and John, Loewenstein, and Prelec (2012) reported evidence that experimental psychologists use some of these tricks. When these tricks are combined, the rejection rate can easily reach more than 50%, even when the null hypothesis is true. As a result, by running improper experiments, it becomes efficient to introduce a file-drawer bias and filter out negative or null findings. Indeed, the natural procedure of someone using data peeking may be to run the experiment until it rejects the null hypothesis or the researcher gives up. In the former case, the experiment is published. In the latter case, the researcher may conclude that there is no effect and choose to not publish the result. Thus, the biases support each other and seriously interfere with the ability of the field to make scientific conclusions about the magnitude of effect sizes.

Does Publication Bias Matter?

An important aspect of the negative impact of publication bias can be realized by considering the three sets of experiments summarized in Table 2. For one of the experiment sets, the null hypothesis was true, but data peeking (with samples starting at 10 and stopping at 30 data points) and a file-drawer bias were used to produce a set of five experiments that all reject the null hypothesis. What is not reported is that 15 additional experiments did not reject the null hypothesis but were not reported. For another experiment set, the true effect size was 0.1. This study picked sample sizes at random between 10 and 30 data points for each group (no data peeking). There were a total of 100 experiments, but only the five experiments that rejected the null hypothesis in a positive direction (experimental group larger than control group) were reported. The other 95 experiments were put in a "file drawer." For a final experiment set, the true effect size was 0.8, and sample sizes were randomly selected to be between 10 and 30 for five experiments. Because of the large true effect size, each of these experiments rejected the null hypothesis. Before reading further, look at the experiment sets in Table 2 and try to determine which set is valid and which set has a true null hypothesis.

At first glance, the three sets of experiments look very similar. All have a range of sample sizes, p values, and effect sizes. Nevertheless, the publication bias test correctly identifies that Sets 1 and 3 have a bias. The experiments in Set 1 had a true null hypothesis, but the experiments were generated with a combined data peeking and file-drawer bias strategy, where only the five experiments that rejected the null hypothesis in a positive direction were reported. For Set 1, the pooled effect size across the reported experiments is $g^* = 0.82$, but because of the small sample sizes, the power of the experiments to

Table 2. Three Sets In Which Every Reported Experiment Rejects the Null Hypothesis.

Set 1				Set 2				Set 3			
$n_1 = n_2$	t	p	g	$n_1 = n_2$	t	p	g	$n_1 = n_2$	t	p	g
10	2.48	.03	1.06	21	2.67	.01	0.81	16	2.10	.04	0.72
28	2.10	.04	0.55	27	4.72	<.01	1.26	19	2.19	.04	0.70
10	3.12	.01	1.34	22	3.66	<.01	1.08	25	2.22	.03	0.62
15	2.25	.04	0.80	26	2.74	.01	0.75	14	2.24	.04	0.82
12	2.34	.03	0.92	24	2.06	.05	0.58	23	2.49	.02	0.72

Note: For one set, the null hypothesis is actually true, but data peeking and a file-drawer bias caused reporting of only the findings that reject the null hypothesis. For one set, the null hypothesis is false, but the true effect size is only 0.1. A file-drawer bias was used to filter out those experiments that did not reject the null hypothesis. For another set, no bias was used and the null hypothesis is false, with a true effect size of 0.8. See the text for details about which experiment set is valid.

detect this effect size is rather small. The expected number of times these five experiments would reject the null hypothesis is the sum of the power values, which is only 2.8. The probability of all five experiments rejecting the null hypothesis is the product of the power values, which is only .042. This set of experiments has too much successful replication to be believable.

For the experiments in Set 3, the true effect size was only 0.1, and the reported experiments were the five experiments out of 100 that rejected the null hypothesis in the positive direction. In contrast to the small true effect size, the pooled effect across these five experiments is $g^* = 0.70$. On the basis of the power of the experiments to detect such an effect size, the expected number of times these experiments would reject the null hypothesis is only 2.8. The probability of all five experiments rejecting the null hypothesis is only .052, which is below the .1 criterion for publication bias tests.

In contrast, for experiment Set 2, the pooled effect size is $g^* = 0.89$ (because of random sampling, it happens to overestimate the true effect size). The power values for these experiments are all above 0.8, and the expected number of times these five experiments should reject the null hypothesis is the sum of the power values, which is 4.3. The probability of all five of these experiments rejecting the null hypothesis is the product of the power values, which is .45. Thus, the experiment set without bias is believable.

Identifying publication bias

There is a quick and dirty way of identifying publication bias in a set of experiments, and it is the basis of the publication bias test proposed by Begg and Mazumdar (1994). It is not as precise as the test of Ioannidis and Trikalinos (2007), but it serves as a pretty good indicator of bias in many cases. Compare the effect size with the sample size across the experiments in Data Set 1. As the sample size increases, the effect size decreases. This happens because an experiment rejects the null hypothesis only when the t statistic is large enough. Since the t statistic is directly related to the effect size and to the square root of the sample size, reported experiments with smaller samples tend to have larger effect sizes (otherwise they would not have rejected the null hypothesis). Pearson's correlation between sample size and effect size for the experiments in Data Set 1 is $r = -.86$. The same kind of relationship holds for the experiments in Set 3, where $r = -.83$. In contrast, the experiments in Set 2 were created without a publication bias, and the effect size is not negatively related to the sample size ($r = .25$). The corresponding correlation for the experiments from Galak and Meyvis (2011) in Table 1 is $r = -.86$ (using the average of the effect size values for Study 1), which is consistent with the analysis above. This approach to identifying bias should be used cautiously because a similar relationship between sample size and effect size will be found if the experiments have different true effect sizes and an a priori power analysis was used to identify an appropriate sample size for each experiment (Renkewitz et al., 2011). Of course, the publication bias test described above should not be applied to experiments with different true effect sizes.

The experiment sets in Table 2 make it clear that publication bias casts substantial doubt about the validity of a set of experiments. A set of experiments where the null hypothesis is true can give the illusion of being a very strong finding, if the experiments are contaminated by publication bias. Likewise, experiments can have publication bias when the true effect size is not zero. For the experiment sets in Table 2, we know the true effect sizes, but in general practice, this is not possible. This means that, given a set of experiments that successfully replicate too often, it is impossible to determine the true magnitude of the effect and even impossible to determine whether the effect investigated by those experiments is real or false.[2] A contaminated experiment set offers no trustworthy information, and a researcher interested in discovering the truth about the tested phenomenon will need to run new experiments without bias. Some approaches try to compensate for publication bias (e.g., Hedges & Olkin, 1985) by estimating the number of unpublished null findings, but these methods deal only with the file-drawer problem and will give misleading conclusions if other kinds of bias are also present. Thus, only a new set of unbiased experiments can determine whether an effect is real.

Conclusion

Although experimental psychologists have long believed that successful replication is a way to demonstrate the validity of an empirical finding, this belief is not always true. When experiments have low or moderate power, there should frequently be experimental findings that fail to replicate a result, even if the effect is true. In such a situation, it is possible to have too much successful replication, which suggests some form of publication bias.

The presence of publication bias causes real harm to scientific investigations of phenomena. As Figure 1 shows, publication bias can overestimate effect sizes and thereby mischaracterize relationships between variables. In practical terms, treatments or methods that appear to produce big effects in biased research experiments will not be as effective when put to practice. Moreover, once publication bias is identified, it is not possible to use that set of experiments to determine whether the true effect is different from zero.

Misunderstandings about the properties of replication may partly explain why some subfields of psychology (and other fields, such as biology and medicine) do not encourage replication attempts. There appears to be a tendency to believe that once an effect has been shown to be statistically significant, then its truth has been established. With such a view, it is pointless to run additional experiments with the same methods because nothing is gained. In reality, every experimental outcome has uncertainty and there is much to be gained by pooling findings across experimental replications. At some point, there are diminished returns for additional replications, but a single finding that just barely reaches statistical significance is poorly established. If such an experiment is repeated with new random samples of the same size, they are expected to reject the null hypothesis only half of the time. Because researchers misunderstand the nature of replication, they are unmotivated to attempt replications, and when such attempts are made, the results are frequently misunderstood.

The extent of publication bias throughout the field is not known, but the publication bias test of Ioannidis and Trikalinos (2007) can be used to detect it. Given that a study such as Galak and Meyvis (2011), which exceeds the standards of the field in many respects, shows strong evidence of publication bias, there is reason to fear that such bias is endemic. The standards of experimental psychology may actually be encouraging publication bias. Addressing these problems will require significant changes to how experimental psychologists draw conclusions from experiments. The first step is to focus on precisely measuring effects rather than on rejecting the null hypothesis. Having too many rejections of the null hypothesis can mislead a researcher, but there is never direct harm in improving the precision of measurement.

Declaration of Conflicting Interests

The author declared no conflicts of interest with respect to the authorship or the publication of this article.

Notes

1. If anything, the .1 criterion is probably too small (Schimmack, in press) because scientists will be reluctant to use or build on scientific results that have a low probability of occurring (say .15). The criterion value is also typically much larger than the true Type I error rate because the test is quite conservative (Francis, 2012b). As for all hypothesis tests, researchers can choose error rates that they are comfortable with, but it would be difficult to justify setting the Type I error rate for a conclusion of bias to be much smaller than the criterion that was used for the experiments that concluded evidence for an effect.
2. Perhaps some kinds of biases can be identified and dealt with by exploring raw data or discussing the nature of bias with authors. However, in many cases, authors may not know the nature of the bias.

References

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

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Berger, J., & Berry, D. (1988). The relevance of stopping rules in statistical inference (with discussion). In S. S. Gupta & J. Berger (Eds.), *Statistical decision theory and related topics* (Vol. 41, pp. 29–72). New York, NY: Springer.

Champely, S. (2009). *pwr: Basic functions for power analysis* (R package version 1.1.1). Retrieved from http://CRAN.R-project.org/package=pwr

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Dutilh, G., Vanderkerckhove, J., Forstmann, B. U., Keullers, E., Brysbaert, M., & Wagenmakers, E.-J. (2012). Testing theories of post-error slowing. *Attention, Perception, & Psychophysics*, *74*, 454–465.

Fisher, R. A. (1956). Mathematics of a lady tasting tea. In J. R. Newman (Ed.), *The world of mathematics*, Vol. 3 (pp. 1512–1521). Mineola, NY: Courier Dover Publications. (Original work published 1935)

Francis, G. (2012a). Evidence that publication bias contaminated studies relating social class and unethical behavior. *Proceedings of the National Academy of Sciences, USA*, *109*, E1587.

Francis, G. (2012b). Response to author: Some clarity about publication bias and wishful seeing. *i-Perception*, *3*. doi:10.1068/i0519ic

Francis, G. (2012c). The same old new look: Publication bias in a study of wishful seeing. *i-Perception*, *3*, 176–178.

Francis, G. (2012d). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151–156.

Francis, G. (in press). Publication bias in "Red, rank, and romance in women viewing men" by Elliot et al. (2010). *Journal of Experimental Psychology: General*.

Galak, J., & Meyvis, T. (2011). The pain was greater if it will happen again: The effect of anticipated continuation on retrospective discomfort. *Journal of Experimental Psychology: General*, *140*, 63–75.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648.

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245–253.

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.

Kruschke, J. K. (2010). *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York, NY: Academic Press/Elsevier Science.

McCullough, B. D., & McWilliams, T. P. (2010). Baseball players with the initial "K" do not strike out more often. *Journal of Applied Statistics*, *6*, 881–891.

McCullough, B. D., & McWilliams, T. P. (2011). Students with the initial "A" don't get better grades. *Journal of Research in Personality*, *45*, 340–343.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (ISBN 3-900051-07-0). Available from http://www.R-project.org/

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

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*(2). Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2012/february-11-2012-observer-publications/psychology';s-woes-and-a-partial-cure-the-value-of-replication.html

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Schimmack, U. (in press). The ironic effect of significant results on the credibility of multiple study articles. *Psychological Methods*.

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.

Sterne, J. A., 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*, 1119–1129.

Strube, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods*, *38*, 24–27.

Yuan, K. H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, *30*, 141–167.