Psychon Bull Rev (2012) 19:975–991
DOI 10.3758/s13423-012-0322-y

THEORETICAL REVIEW

Publication bias and the failure of replication in experimental psychology

Gregory Francis

Published online: 4 October 2012

Abstract Replication of empirical findings plays a fundamental role in science. Among experimental psychologists, successful replication enhances belief in a finding, while a failure to replicate is often interpreted to mean that one of the experiments is flawed. This view is wrong. Because experimental psychology uses statistics, empirical findings should appear with predictable probabilities. In a misguided effort to demonstrate successful replication of empirical findings and avoid failures to replicate, experimental psychologists sometimes report too many positive results. Rather than strengthen confidence in an effect, too much successful replication actually indicates publication bias, which invalidates entire sets of experimental findings. Researchers cannot judge the validity of a set of biased experiments because the experiment set may consist entirely of type I errors. This article shows how an investigation of the effect sizes from reported experiments can test for publication bias by looking for too much successful replication. Simulated experiments demonstrate that the publication bias test is able to discriminate biased experiment sets from unbiased experiment sets, but it is conservative about reporting bias. The test is then applied to several studies of prominent phenomena that highlight how publication bias contaminates some findings in experimental psychology. Additional simulated experiments demonstrate that using Bayesian methods of data analysis can reduce (and in some cases, eliminate) the occurrence of publication bias. Such methods should be part of a systematic process to remove publication bias from experimental psychology and reinstate the important role of replication as a final arbiter of scientific findings.

G. Francis (✉)
Department of Psychological Sciences, Purdue University,
703 Third Street,
West Lafayette, IN 47906, USA
e-mail: gfrancis@purdue.edu
URL: http://www1.psych.purdue.edu/~gfrancis/home.html

Keywords Bayesian methods · Hypothesis testing · Meta-analysis · Publication bias · Replication

Introduction

Imagine that you read about a set of 10 experiments that describe an effect (call it effect "A"). The experiments appear to be conducted properly, and across the set of experiments, the null hypothesis was rejected 9 times out of 10. Next, you read about another set of 19 experiments that describe effect "B." Again, the experiments appear to be conducted properly, and the null hypothesis was rejected 10 times out of 19 experiments. I suspect most experimental psychologists would express stronger belief in effect A than in effect B. After all, effect A is so strong that almost every test was statistically significant, while effect B rejected the null hypothesis only about half of the time. Replication is commonly used to weed out false effects and verify scientific truth. Unfortunately, faith in replication is unfounded, at least as science is frequently practiced in experimental psychology.

Effect A is based on the series of experiments by Bem (2011) that reported evidence of people using Psi ability to gain knowledge from the future. Even after hearing about this study's findings, most psychologists do not believe that people can get information from the future. This persistent disbelief raises serious questions about how people should and do interpret experimental findings in psychology. If researchers remain skeptical of a finding that has a 90% successful replication rate, it would seem that they should be skeptical of almost all findings in experimental psychology. If so, one must consider whether it is worthwhile to spend time and money on experiments that do not change anyone's beliefs.

Effect B is based on a meta-analysis of a set of experiments that describe the bystander effect (Fischer et al., 2011), which is the empirical observation that people are

less likely to help someone in distress if there are other people around. I do not know of anyone who doubts that the bystander effect is real, and it is frequently discussed in introductory psychology textbooks (e.g., Nairne, 2009). Given the poor replicability of this effect, the persistent belief in the phenomenon is curious.

Contrary to its central role in other sciences, it appears that successful replication is sometimes not related to belief about an effect in experimental psychology. A high rate of successful replication is not sufficient to induce belief in an effect (Bem, 2011), nor is a high rate of successful replication necessary for belief (Fischer et al., 2011). The insufficiency of replication is a very serious issue for the practice of science in experimental psychology. The scientific method is supposed to be able to reveal truths about the world, and the reliability of empirical findings is supposed to be the final arbiter of science; but this method does not seem to work in experimental psychology as it is currently practiced.

Counterintuitive as it may seem, I will show that the order of beliefs for effects A and B is rational based largely on the numbers of reported successful replications. In a field like psychology that depends on techniques of null hypothesis significance testing (NHST), there can be too much successful replication, as well as too little. Recognizing this property of the field is central to sorting out which effects should be believed or doubted.

Part of the problem is that replication's ability to sort out the truth from a set of experiments is undermined by publication bias (Hedges & Olkin, 1985; Rosenthal, 1984). This bias may be due to selective reporting of results that are consistent with the desires of a researcher, or it may be due to the desires of editors and reviewers who want to publish only what they believe to be interesting findings. As is shown below, a publication bias can also be introduced by violating the procedures of NHST. Such a publication bias can suggest that a false effect is true and can overestimate the size of true effects. The following sections show how to detect publication bias, provide evidence that it contaminates experimental psychology, and describe how future studies can partly avoid it by adopting Bayesian data analysis methods.

A publication bias test

Ioannidis and Trikalinos (2007) described a test for whether a set of experimental findings contains an excess of statistically significant results. Because this test, which I call the publication bias test, is central to the present discussion, the method will be described in detail.

The ability of repeated experiments to provide compelling evidence for the validity of an effect must consider the statistical power of the experiments. If all of the experiments have high power (the probability of rejecting the null hypothesis when it is false), multiple experiments that reject the null hypothesis will indeed be strong evidence for the validity of an effect. However, the proportion of times a set of experiments rejects the null hypothesis needs to reflect the underlying power of those experiments. Even populations with strong effects should have some experiments that do not reject the null hypothesis. Such null findings should not be interpreted as *failures* to replicate, because if the experiments are run properly and reported fully, such nonsignificant findings are an expected outcome of random sampling. As is shown below, some researchers in experimental psychology appear to misunderstand this fundamental characteristic of their science, and they engage in a misguided effort to publish more successful replications than are believable. If there are not enough null findings in a set of moderately powered experiments, the experiments were either not run properly or not fully reported. If experiments are not run properly or not reported fully, there is no reason to believe the reported effect is real.

The relationship between power and the frequency of significant results has been made many times (e.g., Cohen, 1988; Gelman & Weakliem, 2009; Hedges, 1984; Sterling, Rosenbaum, & Weinkam, 1995), but there has not been a systematic way to test for violations of this relationship. The publication bias test identified by Ioannidis and Trikalinos (2007) uses the reported effect sizes to estimate the power of each experiment and then uses those power measures to predict how often one would expect to reject the null hypothesis. If the number of observed rejections is substantially larger than what was expected, the test indicates evidence for some kind of publication bias. In essence, the test is a check on the internal consistency of the number of reported rejections, the reported effect size, and the power of the tests to detect that effect size.

A difference between the expected and observed numbers of experiments that reject the null hypothesis can be analyzed with a χ^2 test:

$$\chi^2(1) = \frac{(O - E)^2}{E} + \frac{(O - E)^2}{(N - E)}, \quad (1)$$

where O and E refer to the observed and expected number of studies that reject the null hypothesis. N is the total number of studies in the set of reported experiments. This analysis actually tests for both too many and too few observed rejections of the null hypothesis, relative to the expected number of rejections. The latter probabilities will be negligible in most situations considered here.

The observed number of rejections is easily counted as the number of reported experiments that reject the null hypothesis. The expected number of rejections is found by first estimating the effect size of a phenomenon across a set

of experiments. This article will consider only mean-based effect sizes (such as Hedges' g), but a similar analysis could be used for correlation-based effect sizes. Meta-analytic methods (Hedges & Olkin, 1985) weight an experiment-wise effect size by its inverse variance to produce the pooled estimate of the effect size.

With the pooled estimated effect size and the sample size(s) of each experiment, it is easy to determine the power of each experiment to reject the null hypothesis. Power is the complement of type II error, β, (the probability of failing to reject the null hypothesis when the effect size is not zero). For a set of N experiments with type II error values β_i, the expected number of times the set of experiments would reject the null hypothesis is

$$E = \sum_{i=1}^{N} (1 - \beta_i), \tag{2}$$

which simply adds up the power values across all of the experiments.

For relatively small experiment sets, an exact test can be used to compute the probability of getting an observed number of rejections (or more) for a given set of experiments. Imagine a binary vector, $\mathbf{a} = [a_1, \ldots, a_N]$, that indicates whether each of N experiments rejects the null hypothesis (1) or not (0). Then the probability of a particular pattern of rejections and nonrejections can be computed from the power and type II error values:

$$\text{Prob}(\mathbf{a}) = \prod_{i=1}^{N} (1 - \beta_i)^{a_i} \beta_i^{(1-a_i)}. \tag{3}$$

The equation is simply the product of the power and type II error values for the experiments that reject the null hypothesis or fail to reject the null hypothesis, respectively. If every experiment rejects the null hypothesis, the term will simply be the product of all the power values. Likewise, if no experiment rejects the null hypothesis, the term will be the product of all the type II error values.

Fisher's exact test can be used to compute the probability of the observed number of rejections, O, or more rejections. If the vectors that describe the different combinations of experiments are designated by an index, j, then the probability of a set of experiments having O or more rejections out of a set of N experiments is

$$\text{Prob}(O \text{ or more experiments reject}) = \sum_{k=O}^{N} \sum_{j=1}^{{}_N C_k} \text{Prob}(\mathbf{a}(j)), \tag{4}$$

where ${}_N C_k$ indicates N choose k, the number of different combinations of k rejections from a set of N experiments, and j indexes those different combinations. If all of the experiments in a reported set reject the null hypothesis, there is only one term under the summations, and Eq. 4 becomes the product of the power values.

Following the standard logic of hypothesis testing, if the probability of the experiments is small under the hypothesis that there is no bias, there is evidence of publication bias in the set of experiments. It is not entirely clear what is the appropriate criterion of "small" for a publication bias test. Using the traditional .05 criterion seems odd, since the purpose of such a low criterion for NHST is to ensure that researchers do not mistakenly conclude that they have evidence for an effect that does not really exist. In contrast, when the probability is below the criterion for a publication bias test, one concludes that the set of experiments does *not* have evidence for an effect. Thus, the more conservative approach, relative to the conclusion about the existence of an empirical effect, is to use a larger criterion value. Of course, if the criterion is too large, the test will frequently report evidence of publication bias where it does not exist. As a compromise between these competing demands, tests of publication bias frequently use a criterion of .1 (Begg & Mazumdar, 1994; Ioannidis & Trikalinos, 2007; Sterne, Gavaghan, & Egger, 2000). As will be shown below, the publication bias test is quite conservative, so this criterion is often larger than the type I error rate of the test.

The following sections demonstrate how the publication bias test works by looking at simulated experiments. The advantage of starting with simulated experiments is that there is a known ground truth about the presence or absence of publication bias.

File drawer bias

A file drawer bias is based on the idea that researchers more frequently publish experimental findings that reject the null hypothesis (Rosenthal, 1984; Scargle, 2000; Sterling, 1959). The impact of such a bias on the interpretation of a set of experiments and the ability of the publication bias test to identify a file drawer bias were investigated with simulation studies.

In a simulated experiment of a two-sample t-test, a common sample size for each group was chosen randomly to be between 15 and 50. For a control sample, n_1 scores were drawn from a normal distribution with a mean of zero and a standard deviation of one. For an experimental sample, $n_2 = n_1$ scores were drawn from a normal distribution with a mean of 0.3 and a standard deviation of one. A two-sample, two-tailed t-test was then computed for the samples with $\alpha = .05$. The experiment was repeated 20 times with different sample sizes and random samples, and for each experiment, an estimate of effect size was computed as Hedges g with a correction for bias in effect size (Del Re, 2010; Hedges & Olkin, 1985).

Table 1 summarizes the key statistics for this set of experiments. The fourth column gives the estimated value of Hedges' *g* for each experiment. Because of random sampling, it varies around the true effect size. The next three columns give estimates of experimental power (Champely, 2009; R Development Core Team, 2011) for three different effect sizes, which are given below the power values. The first power column is for the true effect size. The second power column is for a meta-analytic pooled effect size across all 20 experiments. This estimated effect size is quite close to the true effect size, so the power values for the true and pooled effect size columns are also quite similar. For each column, the sum of the power values is the expected number of times the null hypothesis would be rejected. Not surprisingly, these values are close to the observed five rejections out of 20 experiments.

The situation is different if the experiments are filtered by a file drawer publication bias. Suppose that the calculations above ignored the null findings (because they were not published) and, instead, were based only on the five experiments that rejected the null hypothesis. Using the effect size pooled from only those experiments, the penultimate column in Table 1 provides their power values. The experiments that reject the null hypothesis tend to have larger *t* values and experiment-wise effect sizes, so their pooled effect size is almost twice the true effect size. As a result, the estimated power of each experiment that rejects the null hypothesis is substantially larger than the true power. Thus, one side effect of a file drawer publication bias is that both the effect size and experimental power are overestimated (Hedges, 1984; Ioannidis, 2008; Lane & Dunlap, 1978). This overestimation undermines the publication bias test, but there is another effect that overcomes this problem. The sum of the overestimated power values is the expected number of times the null hypothesis would be rejected if the experiments were fully reported. Because the experiments were not reported fully, this sum includes only the five experiments that reject the null hypothesis. Since the power values from the unreported experiments are not included, the sum is substantially smaller than for the fully reported cases. In fact, even though the effect size and power values are overestimated, the calculations would lead one to expect that around three of the five experiments should reject the null hypothesis. The probability that five out of five experiments like these would reject the null hypothesis is found by multiplying the power values, and this gives .081, which is below the .1 criterion used for the publication bias test. Thus, the test correctly identifies that something is amiss with the biased set of experiments.

The probability of five rejections out of the 20 experiments can also be computed for the fully reported set of experiments using the true effect size and the pooled effect

Table 1 Statistical summary of 20 simulated experiments to show the properties of a file drawer publication bias

n_1	n_2	t	Effect size	Power from true ES	Power from pooled ES	Power from biased ES	BF_{10}
29	29	0.888	0.230	0.202	0.206		0.282
25	25	1.380	0.384	0.180	0.183		0.490
26	26	1.240	0.339	0.186	0.189		0.411
15	15	0.887	0.315	0.125	0.126		0.366
42	42	0.716	0.155	0.274	0.279		0.212
37	37	1.960	0.451	0.247	0.251		1.005
49	49	−0.447	−0.090	0.312	0.318		0.171
17	17	1.853	0.621	0.136	0.138		1.039
36	36	2.036	0.475	0.241	0.245	0.718	1.159
22	22	1.775	0.526	0.163	0.166		0.869
39	39	1.263	0.283	0.258	0.262		0.360
19	19	3.048	0.968	0.147	0.149	0.444	9.503
18	18	2.065	0.673	0.141	0.143	0.424	1.429
26	26	−1.553	−0.424	0.186	0.189		0.602
38	38	−0.177	−0.040	0.252	0.257		0.177
42	42	2.803	0.606	0.274	0.279	0.784	5.631
21	21	1.923	0.582	0.158	0.160		1.104
40	40	2.415	0.535	0.263	0.268	0.764	2.362
22	22	1.786	0.529	0.163	0.166		0.882
35	35	−0.421	−0.100	0.236	0.240		0.197
Pooled ES				0.3	0.303	0.607	
Expected number of rejections (E)				4.14	4.214	3.135	
Probability of observed (O = 5) or more rejections				0.407	0.417	0.081	

size. As the bottom row of Table 1 shows, the probability of five or more rejections out of the 20 experiments is around .4, which was measured using an exact test for the 1,042,380 different combinations of experiments that could have five or more rejections of the null hypothesis out of these 20 reported experiments.

The example in Table 1 demonstrates a single case where the publication bias test works well. It is still necessary to show that the test can generally discriminate experiment sets with a file drawer bias from experiment sets that do not contain any bias. Simulations validated the publication bias test by repeating the experiments in Table 1, while varying the size of the experiment set being considered and the probability of null results being published.

Designate the size of the experiment set being tested for publication bias as M. Using the same parameters for each experiment as for the findings in Table 1, 1,000 experiment sets of size M were simulated. The publication bias test was then applied to each of the experiment sets. In order to ease the computations, the χ^2 test for publication bias was used instead of the exact test (these different methods were verified to give essentially the same result for a subset of the experiment sets). If the p value for the χ^2 test was less than or equal to .1 and if the number of published experiments that rejected the null hypothesis was larger than expected, the experiment set was classified as having a publication bias. A publication bias was not reported for the relatively rare cases where the χ^2 test found evidence of fewer than expected statistically significant experiments, which tended to happen only when there was no bias against publishing null results. The ideal publication bias test has a low proportion of experiment sets indicating bias when the experiments are fully published (no bias) and a high proportion indicating bias when some (or all) of the experiments that fail to reject the null hypothesis are not published.

The simulations varied the size of the experiment set, M, between 10 and 80 in steps of 5. The simulations also varied the probability, q, of publishing an experiment that did not reject the null hypothesis. When $q = 0$, only the experiments that reject the null hypothesis are published and, thereby, considered in the publication bias test analysis. When $q = 1$, both significant and null findings are fully published, and there is no publication bias. The simulations also considered intermediate probabilities of publication, $q = .125, .25, .5$, which will produce experiment sets with differing amounts of publication bias.

Figure 1 plots the proportion of times that the test reports evidence of publication bias as a function of the number of experiments in the set. The different curves are for the different probabilities of publishing a null finding. The proportion of times the test falsely reported a publication bias when the experiment set actually published all experiments ($q = 1$) was at most .02. The publication bias test does not have many false alarms when the data are fully reported.

For experiment sets with bias, the test becomes ever more accurate as the number of experiments increases. Under an extreme file drawer bias ($q = 0$), only the experiments that reject the null hypothesis are published, and for the parameters used in these simulations, that corresponds to around 2.2 rejections for every 10 experiments. These values are given in parentheses on the x-axis of Fig. 1. As more null findings are published ($q > 0$), the test becomes less likely to report the presence of a publication bias.

The test is very conservative for small experiment sets. The proportion of times the test detects publication bias does

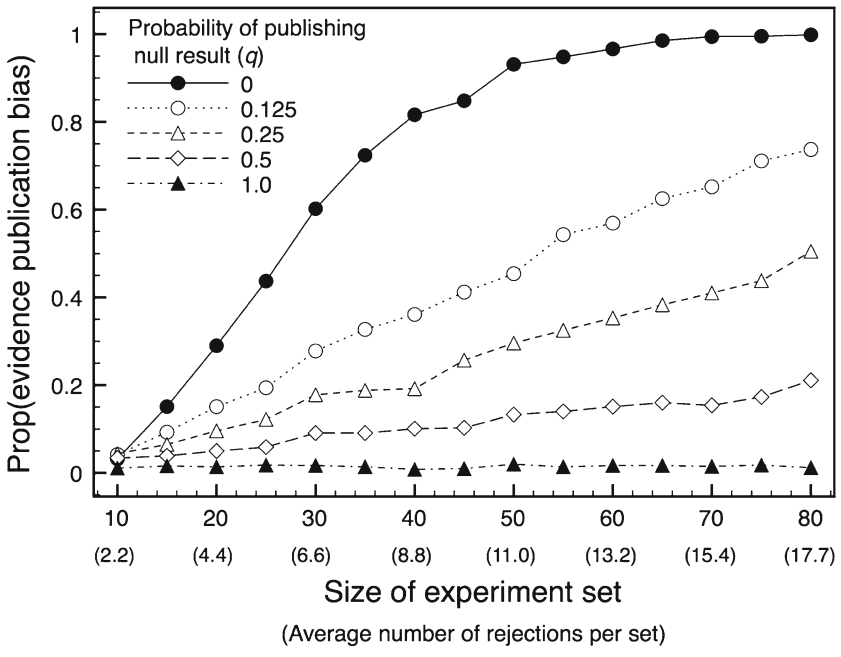

Fig. 1 Results of simulated experiments exploring the ability of the publication bias test to discriminate biased and unbiased experiment sets. When all experiments are published ($q = 1$), the test almost never reports evidence of bias. When only significant experiments are published ($q = 0$), the proportion of times the test reports bias increases with the size of the experiment set. The values in parentheses are the average number of experiments that rejected the null hypothesis in the experiment set

not reach 50% until the experiment set contains between 25 and 30 experiments. This corresponds to an average of between 5.5 and 6.6 reported experiments that reject the null hypothesis. Thus, if an extreme file drawer bias really is present, the test has a fairly low chance of detecting the bias if fewer than 5 experiments are reported. Even so, the low false alarm rate means that when evidence of publication bias is found, the test is very likely to be correct.

It might be tempting to dismiss the simulation findings as being purely theoretical because few researchers would purposely suppress null findings while reporting significant findings. However, there are at least two situations where this kind of behavior may be quite common for researchers who do not understand how methodological choices can produce this kind of bias (see also Simmons, Nelson, & Simonsohn, 2011).

Multiple measures

For some experiments, it is common to measure a wide variety of characteristics for each participant. For some areas of research, there are easily more than 20 such items, and researchers can run multiple statistical analyses but report only the findings that are statistically significant. This type of selective reporting is very similar to a file drawer publication bias. (It is not quite the same, because the measures often have dependencies.) Such an approach violates the basic principles of hypothesis testing and the scientific method, but I have heard more than one academic talk about where it appeared that this kind of selective reporting was being done.

Improper pilot studies

A single reported experiment that rejects the null hypothesis is sometimes the end result of a series of pilot experiments that failed to reject the null hypothesis. Often times, this kind of pilot work is valid exploratory research that leads to a well-designed final study. If such final reported studies have large power values, the publication bias test will not report a problem. On the other hand, some researchers may believe that they are running various pilot studies, but they are really just repeating an experiment with minor variations until they happen to reject the null hypothesis. A set of such experiments with relatively low power will show the pattern exhibited by the biased reports in Table 1 and will likely be detected by the publication bias test.

Data-peeking bias

Traditional hypothesis testing requires a fixed sample size for its inferences to be valid. However, it is easy to violate this requirement and artificially increase the likelihood of rejecting the null hypothesis. Suppose a researcher runs a two-sample *t*-test by gathering data from control and experimental groups with 15 participants. If the difference between groups is statistically significant, the experiment stops, and the researcher reports the result. If the difference between groups is not statistically significant, the researcher gathers data from one more participant in each group and repeats the statistical analysis. This process can be iterated until the difference between groups is found to be statistically different or the experimenter gives up. Concerns about this kind of data peeking (also called optional stopping) have been raised many times (Berger & Berry, 1988; Kruschke, 2010a; Simmons et al., 2011; Strube, 2006; Wagenmakers, 2007), but many researchers seem unaware of how subversive it can be. If the null hypothesis is true, the type I error rate for experiments using this procedure for several iterations can dramatically increase from the expected .05, and it can effectively approach 1.0 if the experimenter is willing to gather enough data. There are many variations of this approach, and the increase in type I error produced by any of these approaches depends on a variety of experimental design factors, such as the number of initial participants, the number of participants added after each peek at the data, and the criteria for continuing or stopping the experiment. All of these approaches introduce a publication bias because they end up rejecting the null hypothesis more frequently than would happen if proper NHST rules were followed. An inflated rejection rate occurs even if the null hypothesis is false.

To explore the effects of publication bias due to data peeking, simulated experiments were generated with a data-peeking method. Each experiment started with a random sample of 15 scores from a control (mean of zero) and an experimental (mean of 0.3) normal distribution, each with a standard deviation of one. A two-tailed *t*-test was performed, and the experiment was stopped; the result was judged significant if $p \leq .05$. If the result was not significant, one additional score from each distribution was added to the data set, and the analysis was run again. This process continued until the sample size reached 50 or the experiment rejected the null hypothesis.

Table 2 shows the statistical properties of 20 such experiments. The sample size values reflect the number of data points at the time the experiment ended. Unlike a file drawer bias, data peeking does not much alter the pooled effect size. For the particular data in Table 2, the pooled estimated effect size from the fully published data set is 0.323, which is only slightly larger than the true effect size of 0.3.

However, the number of experiments that reject the null hypothesis is artificially high under data peeking. The last column of Table 2 lists the power of each experiment for the pooled effect size and the given sample sizes. The sum of these power values is $E = 5.191$, which is much smaller than

Table 2 Statistical summary of simulated experiments (with true effect size equal to 0.3) to show the properties of a data-peeking publication bias

n_1	n_2	t	Effect size	Power from pooled ES (data peeking)	Power from pooled ES (data-peeking and file drawer bias)
37	37	2.009	0.462	0.278	0.759
50	50	1.130	0.224	0.359	
50	50	0.027	0.005	0.359	
39	39	2.009	0.450	0.291	0.781
26	26	2.020	0.552	0.207	0.602
50	50	0.724	0.144	0.359	
19	19	2.419	0.768	0.162	0.469
50	50	0.850	0.169	0.359	
25	25	2.141	0.596	0.201	0.585
50	50	0.716	0.142	0.359	
44	44	2.079	0.439	0.322	0.829
15	15	3.215	1.142	0.137	0.382
17	17	2.197	0.736	0.150	0.426
15	15	2.369	0.842	0.137	0.382
15	15	2.091	0.743	0.137	0.382
19	19	2.191	0.696	0.162	0.469
50	50	0.559	0.111	0.359	
50	50	−0.898	−0.178	0.359	
15	15	3.089	1.097	0.137	0.382
50	50	1.414	0.281	0.359	
Pooled ES				0.323	0.627
Expected number of rejections (E)				5.191	6.447
Probability of observed (O = 12) or more rejections				<.001	<.001

the observed O = 12 significant findings. An exact test (across 263,950 different experiment combinations) reveals that the probability of finding 12 or more experiments that reject the null hypothesis out of 20 experiments with these power values is slightly less than .001. Thus, there is very strong evidence of a publication bias due to data peeking.

The final column of Table 2 shows the estimated power values related to a situation where data peeking was used for each study and a file drawer bias was added so that only the positive findings were reported. As was discussed for the file drawer bias in Table 1, the effect size and power values are substantially overestimated, but the fewer considered experiments means that one would expect only about E = 6.45 of the 12 reported experiments to reject the null hypothesis. The probability that all 12 reported experiments would reject the null hypothesis is the product of the power values, which is only .0003.

Table 3 summarizes a similar simulation analysis when the effect size is zero, which means that the null hypothesis is true. In the simulations, the experiment continued until the null hypothesis was rejected in the right-hand tail (evidence in the negative direction was ignored, under the idea that a researcher was looking for a particular direction of an effect). To be sure to get enough experiments that rejected the null hypothesis, the maximum number of data samples was increased to 100.

Because the null hypothesis was true for these simulated experiments, every rejection of the null hypothesis was a type I error. If the experiments were run properly (without data peeking), 1 out of the 20 experiments would be expected to reject the null hypothesis. In fact, with data peeking, the null hypothesis was rejected four times (type I error rate of 0.2). The pooled effect size is quite small (0.052) because most of the experiments that reject the null hypothesis have small sample sizes, as compared with the experiments that do not reject the null hypothesis, and large samples have more influence on the pooled effect size estimate than do small samples. With such a small effect size, the power values are also quite small, and the expected number of times these 20 experiments will reject the null hypothesis is only 1.284. An exact test considered all 1,047,225 combinations of ways that 4 or more experiments might reject the null hypothesis for these 20 experiments, and the probability of any of those combinations is only .036. Thus, there is clear evidence of publication bias.

The publication bias test continues to properly detect a problem when a file drawer bias is combined with a data-

peeking bias under the situation where the null hypothesis is true. The penultimate column in Table 3 shows the power values for the pooled effect size that was computed from only those experiments that rejected the null hypothesis in a positive direction. As in the previous simulations with a file drawer bias, the effect size and power values are dramatically overestimated. Even so, the expected number of times these experiments should reject the null hypothesis is about half the actual number of reported rejections. The probability that four out of four such experiments would reject the null hypothesis is .047, which indicates publication bias.

It would be good to know how well the publication bias test discriminates between the presence and absence of a data-peeking bias. The performance of the publication bias test should improve as the number of experiments in the set increases and as the effect of peeking introduces more bias. The most biased case of data peeking comes when a single data point is added to each group after each peek at the data. With more data points added after each peek, there are fewer peeks (assuming a fixed upper number of data points), and each addition of a set of data points tends to be less variable. In new simulations, 10,000 simulated experiments like the one in Table 3 (where the null hypothesis is true) were created for each of five data-peeking cases with different numbers of data points added after each peek. From this pool of simulated experiments for a given peek addition number, M experiments were drawn at random, and the set was tested for publication bias (there was no file drawer bias, since experiments were fully reported). The test was repeated 1,000 times, and Fig. 2 plots the proportion of times the test reports evidence of publication bias as a function of the experiment set size, M. The different curves are for variations in the number of added data points after each peek.

The bottom curve (no peeking) verifies that if there is no data peeking (the sample size was always $n_1 = n_2 = 15$), it is rare for the test to report that publication bias was present. The maximum proportion of false alarms was .06 when the experiment set contained only 10 experiments. Just as for the file drawer bias, the publication bias test makes few false alarms.

All of the other curves show results for varying the number of data points added after each peek. The test more frequently detects bias in experiment sets with smaller numbers of data points added after each peek. In the strongest bias case (adding 1 data point for each group after each data peek), the publication bias test detects the bias better

Table 3 Statistical summary of simulated experiments showing the properties of a data-peeking publication bias when the null hypothesis is true

n_1	n_2	t	Effect size	Power from pooled ES (data peeking)	Power from pooled ES (data-peeking and file drawer bias)	BF_{10}
19	19	2.393	0.760	0.053	0.227	2.518
100	100	0.774	0.109	0.066		0.148
100	100	1.008	0.142	0.066		0.181
63	63	2.088	0.370	0.060	0.611	1.065
100	100	0.587	0.083	0.066		0.131
100	100	−1.381	−0.195	0.066		0.278
100	100	−0.481	−0.068	0.066		0.124
100	100	0.359	0.051	0.066		0.118
100	100	−1.777	−0.250	0.066		0.505
100	100	−0.563	−0.079	0.066		0.129
100	100	1.013	0.143	0.066		0.182
100	100	−0.012	−0.002	0.066		0.111
46	46	2.084	0.431	0.057	0.480	1.175
100	100	0.973	0.137	0.066		0.175
100	100	−0.954	−0.134	0.066		0.172
100	100	−0.136	−0.019	0.066		0.112
78	78	2.052	0.327	0.062	0.704	0.920
100	100	−0.289	−0.041	0.066		0.115
100	100	1.579	0.222	0.066		0.368
100	100	0.194	0.027	0.066		0.113
Pooled ES				0.052	0.402	
Expected number of rejections (E)				1.284	2.021	
Probability of observed ($O = 4$) or more rejections				.036	.047	

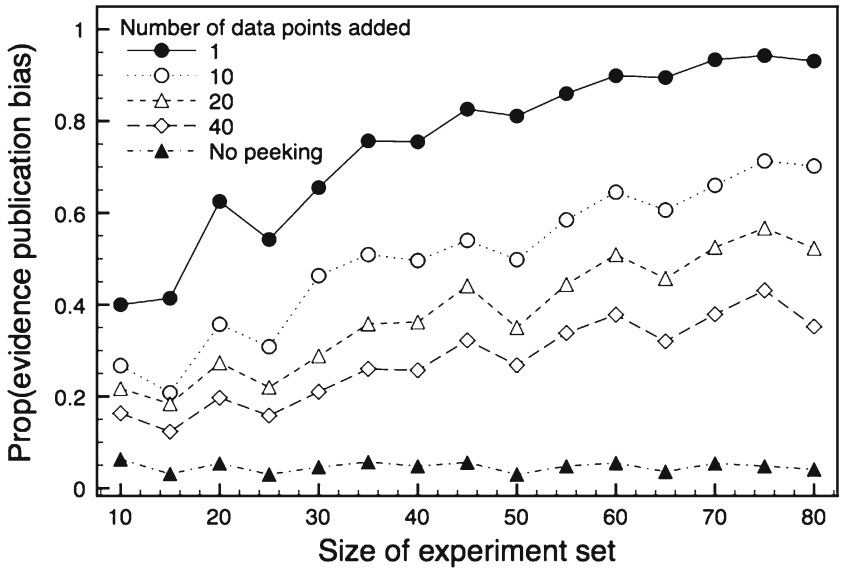

Fig. 2 The results of simulated data-peeking experiments exploring the ability of the publication bias test to discriminate between biased and unbiased experiment sets. When there is no data peeking, the test rarely reports evidence of bias. In the most extreme form of data peeking, the test generally detects the bias more than 50% of the time

than 50% of the time even when the experiment set size is as small as 20. In the least biased case reported here (adding 40 data points to each group after each data peek), there are only three opportunities to test the data, and it is difficult for the test to detect such a bias.

Other sources of publication bias

One of the main strengths of the publication bias test is that it detects biases from a variety of different sources. Bias can be introduced by seemingly innocuous decisions at many different levels of research (Simmons et al., 2011). Given the high frequency of errors in reporting statistical findings (Bakker & Wicherts, 2011), researchers can introduce a publication bias by being very careful when the results are contrary to what was expected but not double-checking results that agree with their expectations. Likewise, data from a participant that performs poorly (relative to the experimenter's hopes) might be discarded if there is some external cause, such as noise in a nearby room; but data from a participant that happens to perform well under the same circumstances might be kept. The key property of the publication bias test is that many of these choices leave evidence of their influence by producing an excess of significant findings, relative to what the estimated effect sizes suggest should be found.

Overall, the publication bias test properly detects evidence for several different types of bias that produce an excess of significant results. One shortcoming of the test is that it is conservative, especially when the number of significant findings is small (less than five), because using the pooled estimated effect size from a set of published experiments gives a generous benefit of the doubt, which often leads to an overestimation of the power values when there really is bias. Thus, once evidence for publication bias is found, the magnitude of the bias is probably larger than what the test directly indicates.

Publication bias in experimental psychology

The previous section described the Ioannidis and Trikalinos (2007) test for publication bias and demonstrated that it works properly for simulated experiments. The test has a low false alarm rate and is conservative about reporting the presence of publication bias. In this section, the test is applied to several sets of studies in experimental psychology. The examples are chosen to highlight different properties of publication bias, but it is also valuable to know that a particular set of experiments has a publication bias, because those findings should be considered nonscientific and anecdotal.

Previous applications of the test have indicated publication bias in individual reports (Francis, 2012a, 2012b, in press) and in a meta-analysis (Renkewitz, Fuchs, & Fiedler, 2011). Schimmack (in press) independently developed a similar analysis. Francis (2012b) found evidence that the precognition studies of Bem (2011), which correspond to effect "A" in the introduction, are contaminated with publication bias. Thus, despite the high replication rate of those studies, the findings appear unbelievable. Below, this conclusion will be contrasted with a similar analysis of effect "B," which had a low replication rate.

It is important to be clear that the indication of publication bias in a set of experiments does not necessarily imply that the reported effect is false. Rather, the conclusion is that the set of experiments does not provide scientific information about the validity of the effect. The proper interpretation of a biased set of experiments is a level of skepticism equivalent to what was held before the experiments were

reported. Future studies (without bias) may be able to demonstrate that the effect is true or false.

Washing away your sins

Zhong and Liljenquist (2006) reported four experiments with evidence that cleansing activities were related to moral purity. For example, they showed that people subject to moral impurity (e.g., recalling unethical behavior in their past) were more likely than controls to complete a word fragment task with cleanliness-related words. Each of the four experiments rejected the null hypothesis, thereby providing converging evidence of a connection between moral purity and cleanliness.

A meta-analytic approach was used to compute the power of the experiments, as summarized in Table 4. For the four experiments reported in Zhong and Liljenquist (2006), the pooled effect size is 0.698, and the penultimate column in Table 4 shows the power of the studies to detect this pooled effect size. The sum of the power values is $E = 2.25$, which is small, as compared with the $O = 4$ rejections. The product of the four power values is the probability that four experiments like these would all reject the null hypothesis if the studies were run properly and reported fully. This probability is .089, which is below the .1 threshold used to conclude evidence of publication bias. Thus, readers should be skeptical about the validity of this work.

Additional positive replications of the effect will alleviate the publication bias only if the effect size of the new experiments is much larger than in the original experiments. One might wonder if a failure to replicate the effect in new experiments might make evidence of publication bias disappear. This would set up a counterintuitive situation where a scientific effect becomes more believable if it *fails* to replicate in new experiments. Mathematically, such a situation seems possible, but it would occur only if the new experiments just barely fail to replicate the effect. The findings of the new experiments contribute to a new meta-analytic estimate of the effect size, and because the experiments do not reject the null hypothesis, their estimate of the effect size will usually be smaller than the previous estimate (assuming similar sample sizes). As a result, the pooled effect size will be smaller, and the power of all the experiments will take smaller values.

In fact, two of the findings of Zhong and Liljenquist (2006) did fail to replicate in a follow-up study (Fayard, Bassi, Bernstein, & Roberts, 2009). The replications reported small effect sizes, as shown in the last two rows of Table 4. Since Fayard et al. used larger sample sizes, their estimates of effect size are weighted more heavily than any individual experiment from Zhong and Liljenquist. Indeed, the pooled effect size across all six experiments is 0.306, which is less than half the previous estimate. As is shown in the final column of Table 4, the power of the experiments to reject the null hypothesis for this new estimated effect size is fairly small. The sum across the power values ($E = 1.62$) is the expected number of times experiments like these six experiments would reject the null hypothesis if they were run properly and reported fully. An exact test considered all possible ways that $O = 4$ or more of the six experiments might reject the null hypothesis and computed the probability of each of the 22 possible combinations. The sum of these probabilities is only 0.036, which means there is still evidence of publication bias in this set of experiments. One cannot draw statistical conclusions from any particular pattern of results, but it is curious that in the reported findings, all of the low-powered experiments rejected the null hypothesis, while the moderately powered experiments failed to reject the null hypothesis.

Bystander effect

One common view of science is that although a given research lab may have biases, these can be mitigated by countering biases from other research labs. Thus, it might seem that although publication bias could be introduced with a few studies for specific topics, the main effects in experimental psychology would likely not be subject to such bias. We can investigate this possibility by looking at a meta-analysis of such a main effect.

Fischer et al. (2011) used meta-analytic techniques to compare the bystander effect for dangerous and nondangerous

Table 4 Statistical properties of the Zhong and Liljenquist (2006) and Fayard, Bassi, Bernstein, and Roberts (2009) experiments on moral impurity and cleanliness. Effect sizes were computed from the reported t and χ^2 tests

Description	*N*1	*N*2	Effect size	Power from pooled ES of Z&L	Power from pooled ES of all
Z&L Experiment1	30	30	0.526	0.757	0.215
Z&L Experiment2	13	13	1.004	0.403	0.117
Z&L Experiment 3	16	16	0.796	0.479	0.142
Z&L Experiment 4	22	23	0.696	0.612	0.178
Fayard et al. Experiment 1	104	106	0.066	NA	0.596
Fayard et al. Experiment 2	57	58	0.228	NA	0.370

situations. The bystander effect is the observation that an individual is less likely to help someone in need if there are other people who might provide assistance. The bystander effect has been widely studied, and the meta-analysis reported 105 independent effect sizes that were based on more than 7,700 participants.

Fischer et al. (2011) broke down the investigations of the bystander effect by several different variables. One important classification was whether the context of the study involved an emergency or a nonemergency situation. (The nonemergency situation is effect "B" from the introduction.) Table 5 shows the results of a publication bias test for each of these two contexts. Given the large number of studies, it is not surprising that evidence for the bystander effect is not always found. Only 24 of the 65 experiments in the emergency situation rejected the null hypothesis (assuming two-tailed *t*-tests with $\alpha = .05$), and only 10 of 19 experiments in the nonemergency situation rejected the null hypothesis. Fischer et al. identified some additional experiments with other contexts that are not considered here.

To know whether there is a publication bias in either of these experiment sets, one computes the power of each experiment for the pooled estimated effect size. For the nonemergency situation, the expected number of experiments that report evidence of the bystander effect (the sum of the power values across experiments) is just a bit less than 11, which is pretty close to the observed number of experiments that reject the null hypothesis. In contrast, the power analysis for the 65 experiments in the emergency situation expects that around 10 of the experiments should find evidence of the bystander effect (most of the experiments have fairly small sample sizes), which is much smaller than the 24 observed findings. The results of the χ^2 tests are shown in the bottom two rows of Table 5.

Even though fewer than half of the studies with an emergency situation reported evidence of the bystander effect, there is strong evidence of a publication bias, so researchers should be skeptical about the validity of this effect. It should be pointed out that this observation is consistent with the general ideas of Fischer et al. (2011), who argued that the bystander effect should be attenuated in emergency situations.

Table 5 Results of the publication bias test for the meta-analyses reported by Fischer et al. (2011) (a negative effect size is evidence for the bystander effect)

	Emergency situations	Nonemergency situations
Number of studies	65	19
Pooled effect size	−0.30	−0.47
Observed number of rejections of H_0 consistent with bystander effect (*O*)	24	10
Expected number of rejections of H_0 consistent with bystander effect (*E*)	10.02	10.77
$\chi^2(1)$	23.05	0.128
p	<.0001	.721

It is noteworthy that there is no evidence of a publication bias for the nonemergency situation studies. The nonemergency situation studies had power values ranging from 0.2 (for several experiments with only 24 participants) to nearly 1.0 (for one experiment with 2,500 participants). The key characteristic of this set of experiments is that the number of studies that reject the null hypothesis is generally consistent with the properties of the experiments and the pooled effect size. This self-consistency is missing in the experiment sets with publication bias.

Altered vision near the hands

The previous two investigations of published data sets were drawn from the subfield of social psychology. There may be different standards and experimental methods for different subfields, but there is no reason to believe that publication bias is a problem only for social psychology. Indeed, data sets that appear to be biased are readily found in mainstream publications for cognitive psychology. For example, Abrams, Davoli, Du, Knapp, and Paull (2008) presented evidence that visual processing is different when a person's hands are near the stimuli. In one representative experiment, they noted that search times increased with display size more quickly when the hands were placed on either side of a computer monitor than when the hands were on the lap. In a set of five experiments, Abrams et al. explored this effect with several different experimental designs and concluded that the effect of hand position is due to an influence on disengagement of attention. Table 6 shows the effect sizes of this phenomenon across the five experiments, which all used a within-subjects design.

The pooled effect size across all five experiments is 0.543. The last column of Table 6 gives the power of each experiment to detect this pooled effect size if the experiments were run properly and reported fully. The sum of the power values across the five experiments is $E = 2.84$, and this expected number of rejections is notably smaller than the reported $O = 5$ rejections. Indeed, the probability that experiments like these would all reject the null hypothesis is the product of the power values, which is .048.

At least one follow-up experiment showed a similar pattern. Davoli and Abrams (2009) reported that the effect was found even with imagined placement of the hands. Their reported result ($n = 16$, $g = 0.567$) is very similar to those in Abrams et al. (2008), but the power of this experiment to detect the pooled effect size is only 0.53. Despite the relatively low power, Davoli and Abrams did reject the

Table 6 Statistical properties of the Abrams, Davoli, Du, Knapp, and Paull (2008) experiments on altered vision near the hands (the effect size for each experiment was computed from the *F* value and the sample size given in the report)

Study	*N*	Effect size	Power from pooled ES
Exp. 1a	20	0.550	0.635
Exp. 1b	12	0.649	0.404
Exp. 1c	20	0.524	0.635
Exp. 2	12	0.652	0.404
Exp. 3	24	0.440	0.722

null hypothesis. If one considers this finding to be another replication of the effect, the probability of six such experiments all rejecting the null hypothesis is roughly .025. Further successful replications of the finding will only provide stronger evidence for publication bias if they include the findings from Abrams et al., so researchers interested in this phenomenon are advised to ignore those findings and start over with new unbiased investigations.

Bayesian data analysis can reduce publication bias

The previous section makes it clear that the publication bias test is not just a theoretical curiosity. There is evidence for publication bias in articles and topics that are published in the top journals in the field and have been investigated by dozens of different labs. It is not yet clear how pervasive the problem might be, but such studies were not difficult to identify. A systematic investigation of publication bias across the field may be necessary in order to weed out the unbelievable findings.

Although no analytical technique is completely immune to all forms of publication bias, this section shows that Bayesian data analysis methods have features that allow them to avoid some forms of publication bias. These properties add to the already excellent motivations to analyze data with Bayesian methods (Deines, 2011; Kruschke, 2010a; Rouder, Speckman, Sun, Morey, and Iverson, 2009; Wagenmakers, 2007).

There are several different approaches to Bayesian data analysis, but in relation to the effects of publication bias, they all have the properties described in this section. The most flexible Bayesian approach involves creating a parametric distribution model for a data set and then using Gibbs sampling techniques to estimate the probabilities of the observed data for different parameters. Readers interested in this approach are encouraged to look at Kruschke (2010b) for an introduction into how experimental psychologists would use these techniques.

Although flexible, the Gibbs sampling approach is computationally intensive. This is usually not a problem for analyzing any particular data set, but for the simulated experiments described below, the computations would be cumbersome. An alternative approach is to accept the assumptions appropriate for a traditional *t*-test and then compute a Bayes factor, which is the ratio of the probability of the observed data for the null and alternative hypotheses. The main advantage of this approach is that Rouder et al. (2009) derived a formula for computing the Bayes factor using a standard objective prior distribution that can be applied to *t*-tests. This formula requires only the sample size(s) and the *t* value that is used in traditional NHST approaches. A complete understanding of the calculations of the Bayes factor is not necessary for the following discussion, but interested readers should see Rouder et al. for details on the *t*-test and Rouder, Morey, Speckman, and Province (in press) for Bayesian analysis methods of ANOVA designs. An additional advantage of using the Bayes factor computations is that Rouder and Morey (2011) derived a meta-analytic version that pools information across multiple experiments. Again, the meta-analytic calculations require only the sample size(s) and *t* value for each experiment.

As a ratio of probabilities given the null and alternative hypotheses, a Bayes factor of one corresponds to equal evidence for both hypotheses. It is arbitrary whether to place the probability corresponding to the null hypothesis in the numerator or denominator of the ratio. I have put the null hypothesis in the denominator and use the term BF_{10} to indicate that this is the Bayes factor with hypothesis 1 (alternative) in the numerator and hypothesis 0 (null) in the denominator. A Bayes factor larger than one indicates evidence for the alternative hypothesis, while a Bayes factor less than one indicates evidence for the null hypothesis.

Bayesian analysts have identified thresholds that are conventionally used to characterize the evidence in favor of the null or alternative hypotheses (Kass & Raftery, 1995). When $BF_{10} > \sqrt{10} \approx 3.16$, there is said to be "substantial" evidence for the alternative hypothesis. Evidence values between 1 and 3.16 are said to be anecdotal or inconclusive. Other category thresholds include above 10 for "strong" evidence and above 100 for "decisive" evidence. A Bayesian analysis can also provide evidence for the null hypothesis, with thresholds set by the inverse of the alternative category boundaries. Thus, when $BF_{10} < 0.316$, there is said to be "substantial" evidence for the null hypothesis. Unlike NHST, none of these threshold categories are sacrosanct, and one frequently finds the "substantial" evidence thresholds rounded to 3 and 1/3. None of these choices will alter the general discussion below.

At the risk of offending my Bayesian colleagues, the discussion below is going to treat these thresholds with more respect than they deserve. Simulated experiments will investigate how file drawer and data-peeking publication biases affect the frequency of finding "substantial" evidence

for the null and alternative hypotheses in a meta-analysis. From the viewpoint of a Bayesian analysis, this focus ignores a lot of useful information. Generally, a Bayesian data analysis is interested in the evidence itself, rather than identifying where the evidence falls within some quasi-arbitrary categories. Nevertheless, I hope to introduce the main issues to the majority of experimental psychologists, who I suspect will look for concepts analogous to "rejecting the null hypothesis," and these Bayesian categories play a somewhat similar role. The main arguments apply equally well to consideration of evidence itself, without the categories.

File drawer bias

The last column of Table 1 shows the results of a Bayesian analysis of the previously investigated set of 20 simulated experiments that were generated with a true effect size of 0.3. Two of the experiments report substantial evidence for the alternative hypothesis, and five experiments report substantial evidence for the null hypothesis. The remaining experiments do not provide substantial evidence for either the null or the alternative hypothesis. However, the meta-analytic Bayes factor that considers all of the experiments finds decisive evidence for the alternative hypothesis (BF_{10} = 44,285).

It might seem that there would be no reason to publish the inconclusive experiments, but they actually provide a lot of information. The meta-analytic Bayes factor based only on the seven experiments with substantial evidence for either of the hypotheses is BF_{10} = 0.534, which is not convincing evidence for either hypothesis but slightly favors the null. Thus, introducing a file drawer publication bias radically alters the Bayesian interpretation of this set of findings. By not reporting the findings from inconclusive experiments, the meta-analysis would fail to recognize the decisive evidence for the effect. Indeed, pooling nonsignificant findings across a large set of experiments in order to extract significant effects is one of the primary motivations for meta-analytic approaches (Hedges & Olkin, 1985).

To explore the effects of publication bias further, the simulated experiment set was repeated 100 times. After gathering data from 20 experiments, the experiment set was subjected to one of four publication bias conditions, described below. The published data from the 20 experiments were then analyzed with the meta-analytic Bayes factor, and the output was classified as substantial evidence for the alternative hypothesis, substantial evidence for the null hypothesis, or inconclusive. The whole process was then repeated for experiments where the effect size equaled zero (the null hypothesis was true). Table 7 reports the number of times the meta-analytic Bayes factor reached each of the decisions.

When there was no publication bias (first-row pair), the meta-analytic Bayesian analysis almost always makes the correct decision, and it never reports substantial evidence for the wrong hypothesis. The second-row pair corresponds to a publication bias where only the experiments that reached a definitive conclusion (substantial evidence for the alternative or null hypothesis) were published. This is the most natural version of a file drawer bias, and it has little effect when the null hypothesis is true. However this bias has a striking negative effect when the true effect size is 0.3. Frequently, the meta-analytically pooled evidence ends up being inconclusive or supportive of the null hypothesis. As was described above, the inconclusive experiments actually contain evidence for the alternative hypothesis, and ignoring that information with a publication bias sometimes leads to the wrong conclusion in the meta-analysis.

The third bias condition supposes that a researcher does not publish any experiment that fails to provide evidence for a positive alternative hypothesis. It might seem that such a bias will lead to artificially enhanced evidence for the alternative hypothesis, and this is indeed the impact if the null hypothesis is true. The impact is not overwhelming, however, and the meta-analytically pooled decision is most often that the evidence is inconclusive. Interestingly, a similar effect is found when the alternative hypothesis is true. As compared with fully publishing all experiments, publishing only those experiments that find evidence for a positive alternative reduces the number of times the meta-analytic Bayes factor data analysis correctly concludes that the alternative hypothesis is true. Again, this is because the unreported experiments contain evidence for the alternative hypothesis and not reporting those findings reduces the ability of the meta-analysis to draw the proper conclusion.

Table 7 The influence of different forms of publication bias on the frequency of each decision of a meta-analytic Bayes factor data analysis for simulated sets of experiments

Type of bias	Meta-analytic decision			
	Effect size	Alternative	Null	Inconclusive
None	0.0	0	95	5
	0.3	100	0	0
Publish only experiments with conclusive evidence (file drawer bias)	0.0	0	100	0
	0.3	50	17	33
Publish only experiments with conclusive positive evidence	0.0	17	0	83
	0.3	90	0	10
File drawer bias based on NHST results (*t*-test)	0.0	23	14	63
	0.3	97	1	2

Finally, the simulation considered the impact of the file drawer bias on the basis of the results of a traditional NHST *t*-test. An experiment was reported only if the *t*-test rejected the null hypothesis. This kind of bias has little impact on the meta-analytic Bayesian analysis if the alternative hypothesis is true, but it greatly diminishes the ability of the meta-analysis to correctly identify when the null hypothesis is true. Instead, the most common decision is that the data are inconclusive.

One main conclusion from the simulations of the meta-analytic Bayes factor data analysis is that a file drawer publication bias undermines the ability of experiments to demonstrate evidence for the alternative hypothesis. Bayesian analysis methods do not reduce the impact of a file drawer bias, but they make the problems with such a bias more obvious. If finding evidence for nonzero effects is of interest to researchers who use a Bayesian analysis, they should avoid publication bias. Moreover, researchers who introduce a publication bias will probably not be very productive, because a common outcome of such bias is to produce experiment sets with inconclusive meta-analysis findings.

Data-peeking bias

A benefit of Bayesian data analysis is that it almost entirely avoids publication bias introduced by data peeking. The last column of Table 3 shows the Bayes factor data analysis for the experiments generated when the null hypothesis was true, but data peeking was used such that the experiment stopped when the null hypothesis was rejected with a positive mean difference. This type of data peeking exaggerates the NHST type I error rate (from 0.05 to 0.2), which leads to a publication bias.

However, Kerridge (1963) proved that the frequency of type I errors with a Bayesian analysis has an upper bound that depends on the criterion used to indicate evidence of an effect. As a result, Bayesian data analysis is mostly insensitive to the presence of data peeking. This insensitivity is evident in Table 3, since not one of the experiments finds substantial evidence for the alternative hypothesis. In fact, 14 out of the 20 experiments (correctly) provide substantial evidence that the null hypothesis is true. When all of the experimental results are considered, $BF_{10} = 0.092$, which is strong evidence for the null hypothesis.

In part, the insensitivity of the Bayesian data analysis to data peeking is due to a stricter criterion for evidence of the alternative hypothesis, relative to NHST approaches. But there are other factors as well. In an additional simulation, data peeking was implemented so that data points were added until the Bayes factor calculation gave a conclusive result for either the null or the alternative hypothesis. Table 8 shows that this approach most often leads to substantial evidence for the null hypothesis, and only one experiment reports substantial evidence for the alternative hypothesis. When pooled across all of the experiments, the meta-analytic Bayes factor is 0.145, which is substantial evidence for the null. An important characteristic of the Bayesian approach, as compared with NHST, is that an experiment can stop with substantial evidence for the alternative hypothesis or for the null hypothesis. Notably, the Bayesian data-peeking approach reaches a conclusion with fairly small sample sizes. Bayesian data analysis fits in well with the general idea that a researcher should gather data until a definitive answer is found.

A Bayesian data analysis is even immune to a data-peeking publication bias that appears to be blatantly deceptive. Table 9 shows experimental data generated with a data-peeking process that continued adding one data point after each peek until the experiment found evidence for a positive alternative hypothesis or the upper limit to the number of data points was reached. In contrast to the data-peeking approach in Table 8, which quickly converged on evidence for the null hypothesis, experiments searching for Bayesian evidence of the positive alternative hypothesis are mostly futile, because the null hypothesis is actually true in these simulations. Eighteen of the 20 experiments reach the upper

Table 8 Statistical summary of simulated experiments showing the properties of a data-peeking publication bias when the null hypothesis is true

n_1	n_2	t	Effect size	BF_{10}
16	16	−0.569	−0.196	0.294
15	15	−0.072	−0.026	0.263
15	15	0.039	0.014	0.263
15	15	0.099	0.035	0.263
24	24	2.696	0.766	4.574
29	29	0.887	0.230	0.282
15	15	0.660	0.234	0.316
15	15	0.018	0.006	0.262
17	17	−0.471	−0.158	0.274
36	36	0.630	0.147	0.215
15	15	−0.085	−0.030	0.263
15	15	−0.534	−0.190	0.296
15	15	0.592	0.210	0.305
15	15	0.589	0.209	0.304
15	15	−0.537	−0.191	0.297
15	15	−0.221	−0.078	0.268
15	15	−0.134	−0.048	0.264
18	18	0.609	0.198	0.286
15	15	0.634	0.225	0.311
15	15	0.404	0.143	0.281

Note. These experiments stopped when the Bayes factor reached a value indicating substantial evidence for either the null or the alternative hypothesis.

limit of the number of data points and then report substantial evidence for the null hypothesis. One experiment stops very early and reports substantial evidence for the positive alternative hypothesis. The remaining experiment reaches the upper limit and reports an inconclusive Bayes factor. The meta-analytic Bayes factor for all experiments gives a value 0.067, which is strong evidence for the null hypothesis.

Finally, Table 10 summarizes the results of simulations that repeated the analyses above and reports on the final decision of the meta-analytic Bayesian analysis under different types of data peeking. The main finding is that data peeking does not introduce a substantial bias for a Bayesian meta-analysis.

Table 10 The influence of different forms of data peeking on the frequency of different decisions of a meta-analytic Bayes factor data analysis for simulated sets of experiments

Type of bias	Meta-analytic decision			
	Effect size	Alternative	Null	Inconclusive
Stop when get conclusive BF_{10}	0.0	0	96	4
	0.3	100	0	0
Stop when get conclusive positive BF_{10}	0.0	0	98	2
	0.3	100	0	0
Stop when reject H_0 for positive t	0.0	0	98	2
	0.3	100	0	0

Conclusions

Science is difficult, and anyone who believes that it is easy to gather good scientific data in a discipline like experimental psychology is probably doing it wrong. The various ways of doing it wrong undermine the ability of replication to verify experimental findings. The publication bias test proposed by Ioannidis and Trikalinos (2007) provides a means of identifying the influence of some of these mistakes, and it should be frequently used as a check on the tendency to report too many significant findings. Experimental psychologists can mitigate the influence of publication bias by using Bayesian data analysis techniques. With these improvements, experimental psychology can fully take advantage of replication to reveal scientific evidence about the world.

Table 9 Statistical summary of simulated experiments showing the properties of a data-peeking publication bias when the null hypothesis is true

n_1	n_2	t	Effect size	BF_{10}
15	15	2.755	1.006	4.918
100	100	−0.519	−0.073	0.126
100	100	−0.299	−0.042	0.116
100	100	0.408	0.058	0.120
100	100	−0.027	−0.004	0.111
100	100	−0.434	−0.061	0.121
100	100	−1.112	−0.157	0.201
100	100	0.643	0.091	0.135
100	100	−0.493	−0.070	0.125
100	100	−1.966	−0.278	0.707
100	100	−0.154	−0.022	0.112
100	100	0.130	0.018	0.112
100	100	−0.660	−0.093	0.137
100	100	−0.563	−0.080	0.129
100	100	−1.466	−0.207	0.312
100	100	0.504	0.071	0.125
100	100	−0.738	−0.104	0.144
100	100	−0.106	−0.015	0.111
100	100	−1.170	−0.165	0.214
100	100	1.011	0.143	0.181

Note. Each experiment stopped adding data points when the Bayes factor showed substantial evidence for a positive alternative hypothesis. Most experiments still report substantial evidence for the null hypothesis.

Some people may mistakenly believe that to avoid a publication bias, researchers must publish the outcome of every experiment. If taken too far, this view leads to the absurd idea that every experiment deserves to be published, even though it may be poorly conceived, designed, or executed. Likewise, some people may believe that all pilot experiments need to be reported to avoid a file drawer bias. Calls for a registry of planned experiments that can be used to check on the outcome of experiments operate along these ideas (e.g., Banks & McDaniel, 2011; Munafò & Flint, 2010; Schooler, 2011). This approach would indeed address issues of publication bias, but at the cost of flooding the field with low-quality experiments. While well intentioned, such registries would probably introduce more trouble than they are worth.

It may seem counterintuitive, but an easier way to avoid a publication bias is to be *more selective* about which experiments to publish, rather than more liberal. The selection criteria must focus on the quality of the scientific investigation, rather than the findings (Greenwald, 1975; Sterling et al., 1995). Studies with unnecessarily small sample sizes should not be published. Pilot studies and exploratory research should not be published as scientific findings. The publication bias test is not going to report a problem with a set of selective experiments as long as the power of those reported experiments is high, relative to the number of reported experiments that reject the null hypothesis.

As was noted above, Bayesian hypothesis testing is better than NHST, but better still is an approach that largely abandons hypothesis testing. Hypothesis testing is good

for making a decision, such as whether to pursue an area of research, so it has a role to play in drawing conclusions about pilot studies. The more important scientific work is to measure an effect (whether in substantive or standardized units) to a desired precision. When an experiment focuses on measurement precision and lets nature determine an effect's magnitude, there is little motivation for publication bias. Better experiments give more precise estimates of effects, and a meta-analysis that pools effects across experiments allows for still more precise estimates. Moreover, one can practice data peeking (e.g., gather data until the confidence interval width for g is less than 0.5) with almost no bias for the final measurement magnitude. Such an approach can use traditional confidence interval construction techniques (Cumming, 2012) or Bayesian equivalents (Kruschke, 2010b). The latter has advantages beyond the issues of publication bias.

The past year has been a difficult one for the field of psychology, with several high-profile cases of fraud. It is appropriate to be outraged when falsehoods are presented as scientific evidence. On the other hand, the large number of scientists who unintentionally introduce bias into their studies (John, Loewenstein, & Prelec, 2012) probably causes more harm than the fraudsters. As scientists, we need to respect, and explicitly describe, the uncertainty in our measurements and our conclusions. When considering publication, authors, reviewers, and editors need to know the difference between good and poor studies and be honest about the work. The publication bias test is available to identify cases where authors and the peer review process make mistakes.

References

Abrams, R. A., Davoli, C. C., Du, F., Knapp, W. H., III, & Paull, D. (2008). Altered vision near the hands. *Cognition, 107,* 1035–1047.

Bakker, M., & Wicherts, J. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43,* 666–678.

Banks, G. C., & McDaniel, M. (2011). The kryptonite of evidence-based I-O psychology. *Industrial and Organizational Psychology, 4,* 40–44.

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, 1* (Vol. 4, pp. 29–72). New York: Springer.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Davoli, C. C., & Abrams, R. A. (2009). Reaching out with the imagination. *Psychological Science, 20*(3), 293–295.

Del Re, A.C. (2010). compute.es: Compute Effect Sizes. R package version 0.2. http://CRAN.R-project.org/web/packages/compute.es/

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290.

Fayard, J. V., Bassi, A. K., Bernstein, D. M., & Roberts, B. W. (2009). Is cleanliness next to godliness? Dispelling old wives' tales: Failure to replicate Zhong and Liljenquist (2006). *Journal of Articles in Support of the Null Hypothesis, 6,* 21–28.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin, 137,* 517–537.

Francis, G. (2012a). The same old New Look: Publication bias in a study of wishful seeing. *Perception, 3*(3), 176–178.

Francis, G. (2012b). 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*.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power. *American Scientist, 97,* 310–316.

Greenwald, A. G. (1975). Consequences of 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., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: 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.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *American Statistical Association, 90,* 773–7935.

Kerridge, D. (1963). Bounds for the frequency of misleading Bayes inferences. *Annals of Mathematical Statistics, 34,* 1109–1110.

Kruschke, J. K. (2010a). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(5), 658–676. doi:10.1002/wcs.72

Kruschke, J. K. (2010b). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press/Elsevier Science.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31,* 107–112.

Munafò, M. R., & Flint, J. (2010). How reliable are scientific studies? *The British Journal of Psychiatry, 197,* 257–258.

Nairne, J. S. (2009). *Psychology* (5th ed.). Belmont, CA: Thomson Learning.

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

Rosenthal, R. (1984). *Applied Social Research Methods Series, Vol. 6. Meta-analytic procedures for social research*. Newbury Park, CA: Sage Publications.

Rouder, J. N., & Morey, R. D. (2011). A Bayes fator meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18,* 682–689.

Rouder, J. N., Morey R. D., Speckman P. L., & Province J. M. (in press). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology.*

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.

Scargle, J. D. (2000). Publication bias: The "File-Drawer" problem in scientific inference. *Journal of Scientific Exploration, 14*(1), 91–106.

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

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470,* 437.

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.

Sterling, T. D. (1959). Publication decisions and the possible effects on inferences drawn from test 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. *The American Statistician, 49,* 108–112.

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.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulleting & Review, 14,* 779–804.

Zhong, C., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science, 313,* 1451–1452.