Follow the argument where it leads: Simonsohn's criticisms on publication bias critiques are unfounded

Gregory Francis
Department of Psychological Sciences
Purdue University
gfrancis@purdue.edu
http://www1.psych.purdue.edu/~gfrancis/pubs.htm
19 July 2012

Simonsohn (2012) argued that my recent publication bias critiques (Francis, 2012a,b,c,d,e,f, in press) are invalid. We both investigate consistency within a set of reported statistical findings to determine whether reported data are valid. Simonsohn recently analyzed cases where authors appeared to be fraudulently manipulating their data (Yong, 2012a,b), while I have focused on cases where publication bias appears to contaminate a set of experimental results (which can happen for a variety of reasons other than fraud).

Simonsohn (2012) describes two criticisms of my approach. (1) My reports of publication bias are themselves biased and thereby invalid. (2) My suggestion to ignore biased data sets is improper and instead researchers should correct for the bias rather than treat them as invalid.

Before turning to specific points below, it is worth noting that Simonsohn's criticisms cannot both be true because they are logically inconsistent. If he really believes that researchers should compensate for the presence of bias (his second criticism), then the existence of bias in my investigations should only lead to a call for a correction, not a claim that the investigations are invalid (his first criticism). This inconsistency is not something that this document will explore further because both of Simonsohn's criticisms are wrong for other reasons.

In the following sections, I rebut both of the criticisms and show that the analyses in my investigations are not only proper but reflect the principles shared by almost all psychological scientists. The headers are those used by Simonsohn (2012).

Biased reporting of publication bias tests

Simonsohn's first criticism is that I have not revealed the number of analyzed cases, so it is impossible to interpret the result of any particular claim that a set of experiments contains publication bias. In particular, it is possible that all such reported findings are Type I errors that report bias when it is not actually present.

The latter statement is true, of course, because we are operating under a situation where the truth is unknown. It is possible that none of the findings I have investigated really have

publication bias but that they just appear to have publication bias. In a similar way, it is possible that all of the positive findings in psychology journals are actually Type I errors. We never know the truth about our statistical decisions.

To emphasize his point, Simonsohn explains how multiple experiments lead to higher possibilities of at least one Type I error. This description is quantitatively correct, but it simply reflects the characteristics of hypothesis testing and uncertainty. To say that the experiment-wise Type I error rate equals 0.05 and that the probability of at least one of 88 such independent experiments making a Type I error equals 0.99 is really just two descriptions of the same fundamental properties of probability and uncertainty.

Along these lines, one of Simonsohn's key concerns is stated on page 4, where he notes, "The question is…whether the conclusion about paper X is false-positive." That is, he wants to know whether a particular finding of publication bias is a Type I error (reporting bias when there really was none). Unfortunately, this is a question that cannot be answered for any hypothesis test. If we knew the answer, we would not need the hypothesis test. Simonsohn wants certainty from a situation where certainty cannot be provided.

What hypothesis testing does provide, whether for an investigation of publication bias or anything else, is control of the rate of making a Type I error. This rate is controlled by the criterion used for the decision to reject the null hypothesis. You can argue that the criterion used in my analyses (0.1) is too lenient or too stringent, but the only way to insure that you never make a Type I error is to never reject the null hypothesis. (Actually, as Francis (2012c,e) noted, the test is very conservative so the true Type I error rate for the publication bias test is often substantially lower than the criterion value. The Type I error rate for many of the published cases is closer to 0.01.)

It is worth noting that the criterion in hypothesis testing defines a conditional probability: the probability of rejecting the null given that it really is true. Thus, the probability of making a false positive decision depends on both the Type I error rate and on the prior probability of the null being true, but the prior probability that a set of experiments is not biased is unknown. So not only do we not know whether a particular decision is a false positive, we do not even know the probability of making a false positive for any particular situation. There is nothing special about publication bias investigations for this conclusion; these difficulties are properties of hypothesis testing in general.

How should a scientist behave in a situation when the truth is unknown? Follow the evidence. If a set of data (or more extreme data) is rare should the null hypothesis be true, then it is justified to tentatively consider the null to be false. This is what psychologists do all the time with regard to their experimental results, and I do the same thing for my investigations of publication bias. If the probability of the experimental outcome is very low without bias (the null is true), then I suggest that we tentatively consider that there is some form of bias (the null is false). There are admittedly difficulties with hypothesis testing, and Bayesian methods might be better in many respects (Kruschke, 2010;

Wagenmakers, 2007), but the basic logic is not entirely unreasonable and the approach is widely practiced.

Now, all of this is not to say that my reported analyses are without bias. As I mentioned to Simonsohn in a phone call, I have looked at several experiment sets with no evidence of publication bias (this, fortunately, includes my own empirical work). I have not (yet) reported any experiment sets that do not show bias (although there is one interesting case that is part of a manuscript currently under review). The presence of bias in my reports might seem like a damning admission, but not all biases invalidate the findings of published results. Invalidation occurs when the bias changes the representation of a phenomenon, but not all biases introduce such misrepresentations.

Suppose I run several experiments to investigate a relationship between afterimages and schizophrenia, and I get a mix of significant and null findings. I can choose to not publish that set of findings (or editors may make this choice for me). Although it introduces a bias, this choice is mostly harmless because it does not invalidate any other findings I might publish on other topics as long as the presence or absence of the unpublished studies does not alter my interpretation of the other topics. A bias to publish findings on some topics and not other topics need not invalidate the properties of the findings that are actually published. (Such selectivity may not be a good scientific strategy, but we are currently concerned only about the statistical validity of published findings.)

What is definitely not harmless is to selectively report significant findings that are all related to the same topic while suppressing null findings on that topic. If I investigate a relationship between afterimages and schizophrenia and get some experiments that reject the null and some experiments that do not, then it is improper for me to publish the significant findings and not the null findings. Such actions would mischaracterize the relationship between afterimages and schizophrenia.

A corollary of this observation is that it may not be useful to consider publication bias for experiment sets that do not address a common phenomenon, such as across a field of study [this is one case where I disagree with Ioannidis & Trikalinos (2007)]. For example, as noted by Sterling (1959) and others, the field of psychology has a preference for publishing statistically significant findings. This selectivity is a bias, and many people interpret this finding to indicate that there are serious problems in psychology (e.g., Bones, 2012). But this kind of bias does not necessarily indicate a problem. A plausible interpretation is that psychologists want to read about (and/or journals want to publish) topics that tend to reject the null hypothesis. Such a bias can exist even if all of the reported findings are themselves unbiased. (I am not arguing that this interpretation is correct, just that it is plausible and consistent with the findings of Sterling.) The way to identify bias is topic-by-topic, because that is where the bias certainly misrepresents the properties of reported findings. This topic-by-topic analysis (sometimes across multiple articles and authors) has been my strategy for exploring publication bias in psychology.

Curiously, the same kind of logic applies to a single experiment set that appears to have bias. For example, Francis (2012b) showed that experiments like those in Balcetis and Dunning (2010) were rather improbable without publication bias. The implication was that the conclusion across the set of experiments in Balcetis and Dunning should not be believed. This interpretation does not require you to believe that each reported experiment is invalid. It is possible that the reported experiments were run and analyzed properly and (individually) give a proper description of the investigated phenomenon. The problem with the findings in Balcetis and Dunning (2010) is not (necessarily) with any individual reported experiment but with the conclusion that is drawn across the set of experiments. This set-wise conclusion can be made invalid when some additional valid experimental findings are missing. With publication bias the experiment set becomes invalid, even if the individual reported experiments are valid.

This logic hinges on the fact that pooling of information across the experiments contributes to a common conclusion, and this is what differentiates the bias in my publication bias investigations from the bias in the findings of Balcetis and Dunning (and others). My investigation of publication bias in Balcetis and Dunning (2010) is unrelated to my investigation of publication bias in Bem (2011), Piff *et al.* (2012), or any other studies. It would not make sense to talk about a conclusion across the sets of experiments that show publication bias because each experiment set investigates a different phenomenon. Each of my investigations stands by itself, so the bias that comes from selectively reporting some cases and not others does not invalidate the reported findings.

Finally, it is worth noting that the bias among my reported investigations is fundamentally different from the bias in the studies I have critiqued. Just because I do not report that a particular experiment set does not appear biased does not prohibit you from analyzing that set. Unlike non-reporting of null experimental findings, where the data are not shared, you are free to explore the properties of any published experiment sets. My unwillingness (or inability) to publish analyses of experiment sets that appear to be without bias does not withhold any data from the field.

Ignore the advice to ignore the data

Simonsohn's second criticism is that my recommendation for how to treat apparently biased experiment sets is unjustified. Indeed, I take a rather strong stand and suggest that biased sets should be considered non-scientific. My recommendation is to run new unbiased experiments to explore the phenomenon of interest. Simonsohn (2012) notes that, in contrast to my rather harsh view, it is common for researchers to try to control for bias rather than flat out reject the data. Such compensatory efforts are common in meta-analyses, but they are fraught with risk.

These efforts always require assumptions about the nature of the bias and then try to do some kind of inverse calculation to produce a "corrected" effect size measurement. My concern is that there is no way to know whether these assumptions are satisfied, and as a result these correction methods provide little confidence (Scargle, 2000). For some forms of

publication bias there is no possible correction, because the reported experimental findings are simply invalid and there is nothing to salvage.

I am not opposed to making corrections when it is possible to identify the source of the bias and compensate appropriately. However, I suspect that such a situation is uncommon for the cases where the publication bias test would indicate a problem. For example, Table 1 lists three sets of five simulated experiments each. Each experiment within a set sampled data from control and experimental normal distributions that differed only in their mean. Thus, within each experiment set the population effect size is held constant (a fixed effect). Every reported experiment rejected the null hypothesis (p<0.05), but only one of the experiment sets was created without bias. Both of the biased experiment sets used a file drawer bias, so there were some additional experiments that did not reject the null hypothesis (or rejected it in the opposite direction), but these were not reported. One of the biased experiment sets also used a data peeking (sometimes called optional stopping) sampling method, where data points were sequentially added to the sample until a significant result was found. Such an approach, although apparently common (John *et al.*, 2012), is a flat out misuse of hypothesis testing because it dramatically increases the rejection rate.

Table 1: Three experiment sets where every reported experiment rejected 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**
20	4.24	<0.01	1.32	22	3.00	<0.01	0.89	25	2.26	0.03	0.63
18	3.17	<0.01	1.03	26	2.36	0.02	0.65	34	2.05	0.05	0.49
26	3.17	<0.01	0.87	13	3.02	0.01	1.15	15	2.24	0.03	0.80
13	2.21	0.04	0.84	28	2.11	0.04	0.56	10	3.07	0.01	1.32
35	3.74	<0.01	0.88	38	2.01	0.05	0.46	25	2.10	0.04	0.59
Pooled effect size			0.97				0.66				0.65

For convenience, the g columns in Table 1 provide a standardized effect size called Hedges g (it is similar to Cohen's d, but corrects for small sample sizes). The bottom row provides the pooled effect size for each experiment set. Given the published experiments, this is the best estimate of the population effect size. However, if the data set is biased, then it is likely that this estimated effect size exaggerates the true population effect size.

The test that I use to detect publication bias (Ioannidis & Trikalinos, 2007) correctly distinguishes the valid experiment set (Set 1, p=0.41) from the biased sets (p=0.07 and p=0.03, respectively). However, this test cannot distinguish different sources of bias. My recommendation for such a situation would be to treat the data in sets 2 and 3 as unscientific, to discard the data, and to run new experiments.

Simonsohn (2012) suggests that instead of discarding the contaminated data sets we should compensate for the bias. He does not specify how to do this, but presumably this compensatory process needs to identify the nature and extent of the bias and then adjust the pooled effect size accordingly. I do not know of any such compensatory process that can meet these needs. The trim-and-fill method is popular, but it often performs quite poorly (Peters *et al.*, 2007), and is not highly recommended (except when alternatives are worse). For the experiment sets in Table 1 this method makes only small adjustments to the pooled effect sizes and it vastly underestimates the number of missing studies. Moreover, the assumptions of the method are not satisfied when data peeking is a source of bias.

I am not aware of any methods that can properly identify and adjust for different kinds of biases. Perhaps an investigation of the raw data might reveal different types of biases and thereby identify some situations where biased data sets can be salvaged. If these methods exist, then it is strange that people do not seem to be using them.

Simonsohn and I discussed some of these issues by email while he was writing his paper. I sent a similar table of experiment sets to Simonsohn by email and asked him to back up his claims by identifying the number of missing studies and the corrected effect size for each experiment set. He never responded, but perhaps the reader can do better. Please perform whatever correction you think is appropriate for sets 2 and 3 in Table 1, and email me your conclusions. If I get some responses, I will post them at my web site. Eventually, I will also reveal the nature of the bias, the true effect sizes, and the number of missing experiments for each set. I should point out that the reader already knows more about these data sets than is typically known about findings in the literature (e.g., the data sets really are biased, the fixed effect model is valid, the types of possible bias are identified).

I do not want to suggest that the experiment sets in Table 1 are typical; the bias in sets 2 and 3 is severe (many null experiments are unpublished). Perhaps adjustment techniques work better when the bias is fairly small, but the test for publication bias test is unlikely to detect the presence of modest bias. Indeed, the test is extremely conservative, so it is only the most egregious situations where it will report the presence of bias. It is exactly those situations where the correction methods are unlikely to work adequately.

For the sake of completeness, consider an example that Simonsohn (2012) uses to discuss the difference between the presence of bias and the significance of bias in a set of experiments. He asks the reader to imagine a set of 100 experiments that each has a power of 0.97 and to imagine that every experiment rejects the null hypothesis. He notes that with the publication bias test, the probability that all 100 experiments would reject the null hypothesis is $(0.97)^{100} = 0.047$, which indicates bias. He then argues that this is a case where one finds evidence of bias (one would expect around three experiments to not reject the null hypothesis) but that the magnitude of the bias is quite small (after all it is only off by three out of 100 experiments).

Simonsohn's example inverts the task researchers face when interpreting a set of experiments. Consider the situation as it would actually be presented to a scientist who does

not have the benefit of knowing that there were really 100 experiments: An area of research has 100 published experiments that all reject the null hypothesis, and each experiment has a power of 0.97 (we will not speculate on how this field of research happened to be so consistent in producing experiments with this power value). To make the situation more concrete consider a set of 100 two-sample *t* tests with a pooled effect size of 0.792 and $n_1=n_2=48$. This gives each experiment an estimated power of 0.97. You run the publication bias test and note that the probability of all 100 experiments rejecting the null is only 0.047. Simonsohn would suggest that you should now try to estimate the magnitude of the bias.

You might want to make an inference about the number of experiments that could produce the data if some null findings were not reported. Unfortunately, we have too many choices because there are two unknowns: the true effect size and the number of unpublished null experiments. Since publication bias tends to produce an overestimate of the true effect size, we should consider effect sizes smaller than that found by the reported experiments. Maybe the true effect size is 0.75, in which case one might expect that there were 5 unpublished null findings out of 105 experiments. But maybe the true effect size is 0.4 and there were 103 unpublished null-findings out of 203 experiments. Or perhaps the true effect size is zero, and questionable research methods increased the Type I error rate to 60% (Simmons, Nelson & Simonsohn, 2011) and there were 67 null findings out of 167 experiments. Since the actual number of experiments is unknown, it is impossible to specify the magnitude of bias in the published set of studies.

Conclusions

It is the responsibility of authors to make a convincing argument that their experiments support their conclusions. When an experiment set appears to be biased it is very difficult to make such an argument, so authors should be highly motivated to avoid bias. Fortunately, avoiding bias is easy: run experiments properly and report all relevant findings, regardless of the statistical outcome.

What appears to be more difficult is to convince people that a statistically significant result actually has a lot of uncertainty (Cumming, 2012). Given the degree of uncertainty in most experimental results, an honest and accurate set of investigations should not always show a statistically significant result, even if the effect is true. This conclusion may be contrary to what many of us were taught in graduate school, but it follows naturally from the mathematics of probability theory.

Given his own interests in identifying fraud (Yong, 2012a,b), I had expected that Simonsohn would appreciate these details, but instead his criticisms of my methods reflect common misunderstandings about hypothesis testing, replication, and statistical inference. Given the clear need to correct these misunderstandings, I hope that Simonsohn will work with me to promote better scientific practice in psychology.

References

Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.

Balcetis, E. & Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science*, **21**(1), 147-152.

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.

Bones, A. K. (2012). We knew the future all along. *Perspectives on Psychological Science*, **7**(3), 307-309. doi: 10.1177/1745691612441216

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

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

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

Francis, G. (2012d). Evidence that publication bias contaminated studies relating social class and unethical behavior. *Proceedings of the National Academy of Sciences USA*, **109**:E1587. doi: 10.1073/pnas.1203591109

Francis, G. (2012e). Checking the counterarguments confirms that publication bias contaminated studies relating social class and unethical behavior. Downloaded from http://www1.psych.purdue.edu/~gfrancis/Publications/FrancisRebuttal2012.pdf

Francis, G. (2012f). Replication initiative: Beware misinterpretation. *Science*, **336**(6083), 802.

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*.

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). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, **1**(5), 658-676. doi:10.1002/wcs.72

Piff, P. K., Stancato, D. M., Côté, S., Mendoza-Denton, R., Keltner, D. (2012). Higher social class predicts increased unethical behavior. *Proceedings of the National Academy of Sciences USA*, **109**, 4086–4091. doi/10.1073/pnas.1118373109

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

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.

Simonsohn, U. (2012). It does not follow: Evaluating the one-off publication bias critiques by Francis (2012a,b,c,d,e,f). Downloaded from http://opim.wharton.upenn.edu/%7Euws/papers/it_does_not_follow.pdf

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.

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

Yong, E. (2012a). The data detective. *Nature*, **487**, 18–19, doi:10.1038/487018a.

Yong, E. (2012b). Uncertainty shrouds psychologist's resignation. *Nature*, doi:10.1038/nature.2012.10968.