Too good to be true: Publication bias in two prominent studies from experimental psychology

Gregory Francis

Department of Psychological Sciences

Purdue University

phone: 765-494-6934

gfrancis@purdue.edu

Word count: 4229

Revised: 24 January 2012

Psychonomic Bulletin & Review (in press)

Abstract

Empirical replication has long been considered the final arbiter of phenomena in science, but replication is undermined when there is evidence for a publication bias. Evidence for publication bias in a set of experiments can be found when the observed number of rejections of the null hypothesis exceeds the expected number of rejections. Application of this test finds evidence of publication bias in two prominent investigations from experimental psychology that purported to reveal evidence of extrasensory perception and to indicate severe limitations in the scientific method. The presence of publication bias suggests that those investigations cannot be taken as proper scientific studies of such phenomena because critical data is not available to the field. Publication bias could partly be avoided if experimental psychologists started using Bayesian data analysis techniques.

Experimental psychologists are trained to use statistics to prevent faulty interpretations of their data. By formalizing the decision process, statistical analysis is supposed to remove the influence of the researcher's belief or desire. No researcher in experimental psychology would report an experiment that involved filtering out subjects who did not behave according to the researcher's expectations because such actions render the findings scientifically meaningless. Publication bias has a similar effect when replication across experiments is used to determine the evidence for experimental findings (Johnson & Yuan, 2007). When replication is the criterion by which scientific results are judged, a bias to publish positive findings is essentially the same error as filtering out subjects who do not behave in a desired way. Even well designed studies can be rendered scientifically useless if other studies are done poorly and publication bias contaminates the set.

We investigated publication bias in two prominent sets of results from experimental psychology. These studies have attracted widespread attention in both academic and non-academic reports because they appear to challenge the established scientific understanding of the universe and the scientific method. Bem (2011) reported empirical evidence that humans can sense future events and use that information to judge the present; an ability that is usually referred to as psi. Convincing evidence for psi would necessitate major alterations in theories of psychology, biology, and physics. Schooler (2011) described the "decline effect," where early empirical investigations show a strong presence of a phenomenon, but later studies show weak or non-existent effects. He speculated that scientific studies might introduce something akin to the Heisenberg uncertainty principle, where observations of effects modify the properties of what is being studied. If this

speculation were true, it would imply a fundamental rethinking of causality and would question the ability of scientific investigations to reveal truths about the world. We report new analyses of the data sets used to support these claims, and we deduce that their conclusions are unwarranted because both sets of data suffer from publication bias. As a result, the studies do not provide useful scientific information about the phenomenon they attempt to study.

Publication bias in Bem (2011)

The psi experiments reported by Bem (2011) have been criticized on both methodological and analytical grounds (Wagenmakers, Wetzels, Borsboom & van der Maas, 2011). However, the methodological criticisms are partly speculative because many experimental steps are not fully described in the published reports. The analytical criticisms are also only partly convincing. Although Wagenmakers *et al* (2011) noted that individual experiments in Bem (2011) do not meet the analytical criteria of a standard Bayesian analysis, Rouder and Morey (2011) used a Bayesian meta-analysis and found some evidence for the proposed psi effect, although these authors emphasized that such evidence must be considered in the context of other conflicting evidence.

Perhaps the most striking property of Bem (2011) is that nine out of ten investigations rejected the null hypothesis, thereby indicating evidence for psi. For many scientists, replication of an effect across multiple experiments provides compelling evidence, but this interpretation is misguided because it does not 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), then multiple experiments that reject the null

hypothesis would indeed be strong evidence for an effect. However, if the experiments have low or moderate power, then even if the effect were real, one would frequently expect to not reject the null hypothesis.

Ioannidis and Trikalinos (2007) used this observation to develop a statistical test for publication bias. The basic idea is to measure the power of each experiment and use those measures to predict how often one would expect to reject the null hypothesis. If the number of reported rejections is substantially different from what was expected, then the test has found 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 sizes, and the power of the tests to detect those effect sizes.

Meta-analytic methods were used to estimate the power of the experiments in Bem (2011), and the key statistical properties of the experiments are shown in Table 1. A pooled effect size was measured across the ten experiments to produce g^{*}=0.186. This pooled effect size differs from the average effect size reported by Bem (2011), because we applied a correction for bias in effect sizes and pooled the effect sizes by weighting the effect size value from each study with its inverse variance (Hedges & Olkin, 1984). This pooled effect size was then combined with the sample size of each experiment to produce an estimated power value (Champely, 2009; R Core Development Team, 2011) for a one-tailed test with α=0.05, which is the hypothesis test Bem used in the original analysis of the data. The penultimate column in Table 1 shows the estimated power of each experiment to detect the pooled effect size. The sum of the power values across the ten experiments is 6.27, which is the expected number of rejections of the null hypothesis given the pooled effect size and

the design of the experiments. The expected number of rejections of the null hypothesis is in stark contrast to the observed nine out of ten rejections.

The probability of getting nine or more rejections for the ten experiments reported by Bem (2011) was calculated with an exact test that computed the probability of every combination of nine or more rejections out of ten experiments by multiplying the appropriate power or Type II error values. There are eleven ways to get nine or more rejections out of ten experiments, and given the estimated powers of these experiments, the probability of getting a set of experiments with nine or more rejections by chance is 0.058, which is less than the 0.1 criterion frequently used for tests of publication bias (Begg & Mazumdar, 1994; Ioannidis & Trikalinos, 2007; Stern & Egger, 2001). This low probability suggests that the reported number of rejections of the null hypothesis is abnormally high given the power of the experiments to detect the pooled effect size.

The use of a pooled effect size supposes a fixed common effect across the experiments, and this approach is consistent with previous interpretations of these experiments (Bem, 2011; Bem, Utts & Johnson, 2011). It is worthwhile to consider the possibility that such pooling is not appropriate, and that each experiment has a different effect size. Such a calculation is frequently called observed power, and although it is a biased estimate of true power (Yuan & Maxwell, 2005), with the large sample sizes used in these experiments it should produce a good estimate of true power, at least on average. These values are given in the last column of Table 1.

The sum of the observed power estimates across the ten experiments is 6.64. The exact test reveals that the probability of getting nine or more rejections by chance from ten

experiments with these power values is 0.088. Again the number of reported rejections of the null hypothesis (evidence for psi) in this set of experiments is higher than is to be expected for the properties of the tests and the magnitude of the effect sizes being measured.

Publication bias in a set of studies showing a decline effect

Schooler (2011) was motivated to explore the decline effect because of its purported influence on reports of verbal overshadowing (Schooler & Engstler-Schooler, 1990). In verbal overshadowing, performance on a visual memory task is impaired after subjects provide a verbal description of the stimuli. The verbal overshadowing effect has a variable history, with some early experiments showing a strong effect and other later experiments showing no effect or an effect in the opposite direction. This weakening of an experimental effect has been labelled the decline effect, and it has been observed for other topics that depend on hypothesis testing, including studies of extrasensory perception. In addition to arguing that scientists should reconsider the fundamental nature of investigating the universe, Schooler (2011) suggested that the decline effect may be quite common but remains hidden because of publication bias. On the other hand, if publication bias is found to contribute to findings with a decline effect, then one must be skeptical about the conclusions drawn about the decline effect itself.

We applied the publication bias test to the set of published experiments identified in a meta-analysis of verbal overshadowing (Meissner & Brigham, 2001). Nine of the eighteen experiments reported evidence of verbal overshadowing (rejected the null hypothesis in a direction consistent with the effect). The pooled effect size (twice the

difference of the arcsine square root proportions as in Cohen, 1988) across all experiments (h^{*} = 0.301) and the sample sizes of each experiment were used to compute each study's power for showing evidence of verbal overshadowing with a two-tailed test. These values are shown in the penultimate column of Table 2. The sum of the power values across all the published experiments was 4.65, which is the expected number of times these studies would report evidence of verbal overshadowing. An exact test computed the probability of each of the 155,382 possible ways to have nine or more of these experiments report evidence for verbal overshadowing. The sum of these probabilities is 0.022, which is the chance probability that nine or more of these experiments would find evidence of verbal overshadowing. The conclusion is that there is a publication bias in these studies that favors reporting evidence of verbal overshadowing. This appears to be true even though only half of the published reports actually found evidence of the effect. A corollary of this analysis is that these experimental studies of verbal overshadowing are woefully underpowered. To determine whether the effect is real, investigators need to run studies with larger sample sizes.

Because the studies of verbal overshadowing tend to use relatively small sample sizes, it is not possible to estimate the power of each experiment with an observed power analysis. Thus, one possible criticism of the publication bias test is that it pooled together findings from experiments that actually investigated different effects. Indeed, the publication bias test is sensitive to heterogeneity of the effect sizes (Ioannidis & Trikalinos, 2007; Johnson & Yuan, 2007). The above analysis addressed this concern by using a selection of experiments that were identified by subject matter experts as attempted replications of the verbal overshadowing effect, but it could be that other experts would

make different choices and thereby lead to different outcomes of the publication bias test. If enough experiments with similar methods were available (e.g., experiments that use a particular set of instructions, or experiments from one laboratory) it would be possible to run the publication bias test for subsets of experiment sets and then pool them together to get an overall probability of the entire set.

Concerns about heterogeneity of effect sizes are often not as worrisome as one might suspect. For example, a reviewer noted that there are two discrepant findings in the studies of verbal overshadowing that show a strong effect in the opposite direction of what is typically reported. These findings could be the result of normal variation due to random sampling from a population with a small effect size (this is the interpretation for the above analysis), but these findings could alternatively be interpreted as investigations of an entirely different effect. If the latter interpretation were true, then the analysis should remove these experiments from the meta-analysis. When this is done, the pooled effect size increases to h^{*}= 0.373 (the two negative experiments had fairly small sample sizes, so they do not strongly influence the pooled effect size). As the final column in Table 2 shows, the power of each experiment increases when considering this larger effect size.

However, the impact of the larger power values for the publication bias test is partly mitigated by the fact that one must now consider the probability of rejecting the null hypothesis nine times out of only sixteen experiments. The sum of the power values for the new effect size is 5.98, and an exact test that considers the 26,333 ways that there could be nine or more experiments that reject the null hypothesis out of these sixteen experiments is 0.090, which is below the 0.1 threshold. Thus, for this data set, even if one uses the

outcome of the experiments to determine whether an experiment is a replication, there is still evidence of publication bias. In general though, using an experimental outcome to determine whether an experiment is an attempted replication is itself a type of bias and should be avoided. Ultimately, the choices about which experiments are replications of a basic effect should be defined by the methodological characteristics of the experiments or by a theoretical explanation of the phenomenon.

Conclusions

Just as no experimental psychologist would believe the findings of an experiment that was biased to report data from only subjects that show a desired effect, so the presence of a publication bias means that the studies of psi and verbal overshadowing do not tell us anything scientifically useful about the phenomena because critical data is not part of the results. The effects may be real, or they may be entirely due to bias. The set of studies are simply not scientific. Even worse, the publication bias test generally cannot identify which, if any, specific experimental results are valid because it only tracks statistics across the entire set. Thus, although some researchers may report their experimental findings fully and properly, those experiments can be rendered scientifically useless by poor reporting practices from other researchers.

It might be feared that the publication bias test is so stringent that almost every set of studies would demonstrate a publication bias. In fact, the test used here is strongly inclined to not report a publication bias because reported effect sizes tend to overestimate reality (Ioannidis, 2008). Moreover, when the publication bias test indicates suppression of null or negative findings, the true effect size for the phenomena being studied is probably

smaller than what is estimated by the biased set of reports. Thus, many of the estimated power values that form the basis of the test are larger than they should be, which means the expected number of rejections is overestimated. Once evidence for bias is found, it is likely that it is even more pronounced than the test indicates.

When evidence of a publication bias is presented, many people think of the file drawer problem, which refers to the idea that a researcher runs many different experiments but only publishes the ones that tend to show evidence for an effect. This kind of bias could be due to the deliberate intention of the author to mislead the field or by an inability to get null findings approved by reviewers and editors. Such a problem surely exists, and the test described above can detect its influence.

A closely related bias, with a similar result, occurs when an experimenter to measures many different variables but then selectively reports only the findings that reject the null hypothesis. Simmons, Nelson & Simonsohn (2011) demonstrated how such an approach (in combination with some other tricks) can suggest evidence for truly outlandish effects. Moreover, seemingly innocuous decisions at many different levels of research can produce a publication bias. Given the high frequency of errors in reporting statistical findings (Bakker & Wicherts, 2011; Wicherts, Bakker & Molenaar, 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 beliefs or expectations. Likewise, data from a subject that happens to perform poorly (given the experimenter's hopes) might be thrown out if there is some external cause, such as noise in a neighbouring room or a mistake in the computer program, but data from a subject that happens to

perform well under the same circumstances might be kept (and even interpreted as evidence of the effect's strength despite distractions).

One form of publication bias is related to experiment stopping rules. When gathering experimental data it is easy to compute statistical tests at various times and explore whether the experiment is likely to work. Such checks (sometimes called data peeking) introduce a bias if the experimenter allows that information to influence whether to continue the experiment (Berger & Berry, 1988; Kruschke, 2010; Wagenmakers, 2007). In a similar way when an experiment ends but the data does not quite reach statistical significance, it is common for researchers to add more subjects in order to get a definitive answer. This approach (although it seems like good science to gather more data) is inconsistent with the traditional frequentist approach to hypothesis testing. Such approaches inflate the Type I error rate (Strube, 2006) and often overestimate effect size.

The publication bias test cannot distinguish between the myriad ways for bias to appear, but since it provides evidence that the studies of psi and verbal overshadowing contain bias, one need not propose radical characteristics of the universe (Bem, 2011) or limits to the scientific method (Schooler, 2011) in order to explain the properties of those studies. The simpler explanation is that, as a set, either the studies are not reporting all relevant information, or the studies rejected the null hypothesis more frequently than they should have because they were run improperly. Either way, these studies are at best anecdotal, and as such they need no scientific explanation at all.

Perhaps the most striking characteristic of both the Bem (2011) study and the set of studies reporting verbal overshadowing is that they meet the current standards of

experimental psychology. The implication is that it is the standards and practices of the field that are not operating properly. Clearly, these standards and practices need to be improved to insure that the frequency of reporting the null hypothesis is consistent with the power of the experiments. Such improvements are long overdue. Sterling (1959) noted that 97.3% of statistical tests rejected the null hypothesis for the major scientific findings reported in four psychology journals. A follow-up analysis by Sterling, Rosenbaum and Weinkam (1995) found a similar result with a null hypothesis rejection rate of 95.56%. These high percentages suggest that the power of the experiments (if the alternative hypothesis is true) must generally be well above 0.9, even though power values in the range of 0.25 to 0.85 are more common in psychology (Hedges, 1984). Formalizing this discrepancy between the observed and expected proportion of rejections of the null hypothesis is the core of the publication bias test developed by Ioannidis and Trikalinos (2007) and used above.

When publication bias has been reported in other fields (e.g., Munafò & Flint, 2010), there is often a call to create a registry of planned experiments and require researchers to describe the outcome of the experiments regardless of the findings (Schooler, 2011). Such a project would be an expensive undertaking, would require complex paperwork for every experiment, and would be difficult to enforce for a field like experimental psychology where many of the investigations are exploratory rather than planned. There is a much simpler partial solution: Bayesian data analysis.

A Bayesian approach has two features that mitigate the appearance of publication bias. First, in addition to finding evidence in support of an alternative hypothesis, a Bayesian analysis can find evidence in support of a null hypothesis. Thus, an experiment

that finds convincing evidence for a null hypothesis provides publishable scientific information about a phenomena in a way that a failure to reject the null hypothesis does not provide. Second, a Bayesian analysis is largely insensitive to the effects of data peeking and multiple tests (Berger & Berry, 1988; Krushke, 2010; Wagenmakers, 2007), so some of the methodological approaches that inflate Type I error rates and introduce a publication bias will be rendered inert. Bayesian data analysis is only a partial solution because there may still be a file drawer problem for researchers, reviewers, and editors who deliberately or unintentionally choose to suppress experimental findings that go against their hopes. Such remaining cases can be identified by the publication bias test described in this paper, or a Bayesian equivalent.

The findings of the publication bias test in experimental psychology demonstrate that the care that is so rigorously taken at the level of an experiment is sometimes not being exercised at the next level up when research findings are being reported. Such publication bias is inappropriate for an objective science, and the field must improve its methods and reporting practices to avoid it.

References

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

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.

Bem, D. J. , Utts, J. & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology,* **101**, 716-719.

Berger, J. & Berry, D. (1988). The relevance of stopping rules in statistical inference (with discussion). In *Statistical Decision Theory and Related Topics 4* (S. S. Gupta and J Berger, eds.), **1,** 29-72. Springer, New York.

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.

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

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

Johnson, V. & Yuan Y. (2007). Comments on 'An exploratory test for an excess of significant findings' by JPA Ioannidis and TA Trikalinos. *Clinical Trials* **4**, 254-255.

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

Meissner, C. A. & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology,* **15**, 603-616.

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

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/.

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

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

Schooler, J. W. & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology,* **22**, 36-71.

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.

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**, 1119-29.

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.

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.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, **100**, 426–432.

Wicherts, J. M., Bakker, M. & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One*, **6**(11): e26828. doi:10.1371/journal.pone.0026828

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

Table 1: Statistical properties of the Bem (2011) experiments on psi. A positive effect size is consistent with psi.

Experiment	Sample size	Effect size	Power from pooled ES	Observed power
1	100	0.249	0.578	0.796
2	150	0.194	0.731	0.765
3	97	0.248	0.567	0.783
4	99	0.202	0.575	0.639
5	100	0.221	0.578	0.710
6a	150	0.146	0.731	0.555
6b	150	0.144	0.731	0.543
7	200	0.092	0.834	0.365
8	100	0.191	0.578	0.598
9	50	0.412	0.363	0.890

Table 2: Statistical properties of the experiments on verbal overshadowing (Meissner & Brigham, 2001). A positive effect size is consistent with verbal overshadowing.

n_1	n_2	Effect size	Power from pooled ES	Power for pooled positive ESs
39	39	0.453	0.264	0.377
33	34	0.498	0.233	0.332
28	28	0.644	0.202	0.286
44	44	0.526	0.292	0.416
35	35	0.692	0.242	0.344
23	27	-0.675	0.184	--
80	80	0.102	0.478	0.654
60	60	0.479	0.378	0.532
30	30	0.175	0.214	0.303
30	30	0.182	0.214	0.303
20	20	0.339	0.157	0.218
40	40	0.245	0.270	0.385
27	32	0.320	0.210	0.297
97	68	0.307	0.478	0.654
25	29	0.659	0.196	0.276
30	30	-0.403	0.214	--
30	30	0.000	0.214	0.303
30	30	0.711	0.214	0.303