Some clarity about publication bias and wishful seeing

Gregory Francis
Department of Psychological Sciences
Purdue University
gfrancis@purdue.edu

Francis (2012) concluded that the findings on wishful seeing described in Balcetis and Dunning (2010) appeared to contain publication bias. In a reply, Balcetis and Dunning (2012), henceforth B&D's response, raised a number of interesting observations and claims. We agree on some points and disagree on other points, and there are several issues that remain unclear. So that it can end on a positive note, this rebuttal will start with the disagreements and finish with the agreements.

Areas of disagreement

B&D's response claims Francis (2012) made a false-positive error

Francis (2012) concluded publication bias in Balcetis and Dunning (2010) because the estimated probability that all five experiments would reject the null hypothesis was less than 0.1. Given the use of a seemingly similar criterion for traditional hypothesis testing, it might appear that there would be a false positive rate of 0.1, but the actual false positive rate for the test is much lower.

To investigate false positives for the publication bias test, I ran a Monte Carlo simulation of experiment sets that were similar to the experiments in Balcetis and Dunning (2011, 2012). Each set consisted of eight experiments that were analyzed with a two-sample, two-tailed *t*-test using a criterion of $p<.05$. The true effect size for each experiment was 0.53 (with samples drawn from normal distributions), and sample sizes were drawn uniformly between 20 and 60 and were equal for both groups. With these experimental properties, the average number of experiments that rejected the null hypothesis was 5.1.

For each set of eight experiments, the publication bias test computed the estimated probability of the findings when all of the experiments were reported (no publication bias). The experiment set was simulated one hundred thousand times and the publication bias test concluded bias 895 times. Thus, the false positive rate was 0.00895. Of course no test that makes decisions under uncertainty can entirely avoid false positives.

Importantly, B&D's response appears to remove some uncertainty. Although the title of B&D's response suggests that Francis (2012) made a false positive error, if their report of a null finding that should have been included in the experiment set is valid, then this suggestion is impossible. The presence of the unpublished null finding means that there was publication bias in Balcetis and Dunning (2010); so rather than making a false positive error, the analysis in Francis (2012) scored a hit.

B&D's response claims Francis (2012) cherry picked cases

In their conclusions, B&D's response suggested that my investigations of publication bias engage in the very practice that I criticize. I would be susceptible to this criticism if I were making inferences about publication bias for the field in general. If a researcher wanted to estimate the frequency of publication bias across a field of study, then it would be critical to take random sets of experiments. Only with a random sample can one validly infer back to the field.

However, the analysis reported in Francis (2012) is similar to a case study report. By definition, such a report is selective and the findings should not generally be extrapolated to situations outside of the particular report. In no way does this selectivity undermine the conclusions in Francis (2012) as applied to the findings of Balcetis and Dunning (2010). Readers are strongly cautioned to not make general inferences about publication bias in the field from this one selective analysis.

Was the finding in Balcetis and Dunning (2010) false?

B&D's response suggested that Francis (2012) made an "egregious error" by inferring that the null hypothesis in Balcetis and Dunning (2010) was true, but I made no such claim. The conclusion of the publication bias analysis was that the experiments in Balcetis and Dunning (2010) did not provide proper evidence for the claimed finding. This conclusion is agnostic on whether the claimed finding is actually true or false.

Even after hearing about evidence for publication bias, researchers may continue to believe that the findings in Balcetis and Dunning (2010) are true because they respect the capabilities of the researchers, because they know of other data, or because they think the basic ideas are sound; but these are beliefs based on information outside of the statistical data reported in Balcetis and Dunning (2010).

Issues that are unclear

Did B&D's response provide "full disclosure"?

In a commendable display of openness, B&D's response explained that there was a suppressed null finding for the work described in Balcetis and Dunning (2010). This is an important admission that validates the main claim in Francis (2012). Although presented as "full disclosure," B&D's response did not address some related issues.

As Francis (2012) noted, there are two broad ways to introduce a publication bias. B&D's response focused on what is commonly called the file-drawer problem, where null findings are suppressed. It is to their credit that B&D's response shared that a study in Balcetis and Dunning (2010) was subject to a file-drawer bias.

However, Francis (2012) noted that a second potential source of publication bias is that the experiments were run improperly and so rejected the null hypothesis more frequently than they should. B&D's response is silent on this issue, and their reported experience with Exp. 3b is difficult to understand. They explain that the reviewers and editor required a new version of this experiment, so they made the requested modifications and re-ran the experiment. It is curious that the published Exp. 3b used fewer subjects than the suppressed Exp. 3b. If Balcetis and Dunning (2010) had used the effect size from the suppressed Exp. 3b to guide the design of the new version (with, say, a power of 0.8), then they should have run an experiment with about 70 participants in each group. Instead, they found a significant result with 26 participants per group (a design that has a power of only 0.4 for the effect size estimated in the suppressed Exp. 3b).

I do not want to read too much into the silence in B&D's response on this issue because it could be that they felt it was unnecessary to state that they did not engage in improper techniques such as optional stopping (checking the data intermittently and stopping the experiment when the null hypothesis was rejected). Likewise, there could have been other issues that guided the design of the reported Exp. 3b. Nevertheless, their disclosure would have been more convincing if it had included a statement categorically denying these kinds of activities.

Publication bias leads to an overestimation of effect sizes

We all agree that publication bias generally leads to overestimation of the true effect size. B&D's response implies that this overestimation is a minor transgression. I agree that if the suppressed null finding is the extent of the bias, then the impact appears to be relatively small. However, the impact need not always be small, and without knowledge about the suppressed experiment, there is no basis for estimating the impact. In particular, overestimation could mean that the true effect size is zero.

Uncertainty about the overestimation of effect sizes is why researchers should go to great lengths to avoid publication bias. An effect may be large and important, but if the experiments that demonstrate the effect contain publication bias, then readers remain uncertain whether the effect is true. It is the responsibility of researchers to demonstrate evidence for their claim, and publication bias makes it very difficult to provide such evidence.

How should effect sizes be pooled? Is the test robust?

B&D's response suggested that the analysis in Francis (2012) is invalid because it pooled effect sizes that should have been treated separately. Francis (2012) also considered this possibility so we agree that it is an issue to be treated seriously. The crux of the question is whether the reported experiments draw samples from populations with a fixed effect size.

It might be tempting to look at the reported effect sizes and use their values to judge whether the experiments have a fixed effect size. Figure 1 shows the effect size and 95%

confidence interval (Kelley, 2007) for each experiment in Balcetis and Dunning (2010) and B&D's response. It may be surprising that the confidence intervals are so large, but if an experiment just barely rejects the null hypothesis (e.g., Exps. 1 and 2b in Balcetis and Dunning (2010)), then it is necessarily true that the range of the confidence interval around the effect size is roughly twice the magnitude of the effect size value. Only Exp. 3a has an effect size that differs much from the other experiments. B&D's response is silent about why Exp. 3a has such a large effect size, even though Francis (2012) specifically noted that it seemed odd[1]. At any rate, all of the confidence intervals show substantial overlap, so there is not much reason to believe that the population effect sizes are different.

More generally, it is difficult to make a statistical argument that experiment effect sizes do, or do not, come from samples taken from a common population. These judgments are especially difficult because the existence of publication bias can introduce heterogeneity among reported effect sizes, even if the true effect size is fixed. An interpretation about the source of effect sizes is almost always driven by a theory or by methodological details.

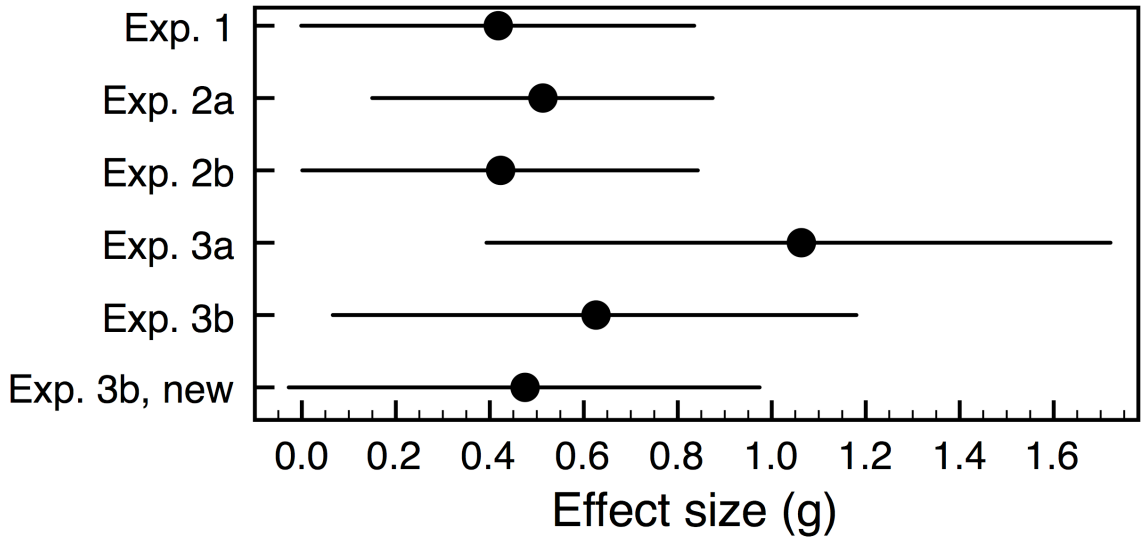

Figure 1. Effect size estimates and 95% confidence intervals for the experiments reported in Balcetis and Dunning (2010, 2012).

Francis (2012) was upfront about the assumptions underlying the choice to pool the effect sizes. B&D's response correctly notes that it is reasonable to challenge these assumptions and to consider alternatives. Although there are surely many more possibilities, they considered two alternative approaches: 1) computing post hoc power for each experiment by using the experiment's effect size; 2) pooling the effect sizes with equal weight.

[1] A text error in Francis (2012) muddled this issue. The last sentence in section 3 should have been (italics indicate the change) "Based on this analysis, one might expect that experiment 3a will have a smaller *effect size* than the other experiments, but Table 1 shows that experiment 3a has the largest effect size of all of the experiments."

Technique 1) is not a good idea unless the sample sizes are large. With small sample sizes and no publication bias, there tends to be much variability in effect size estimates, and the resulting post hoc power values tend to underestimate true power values that are bigger than one half (Yuan & Maxwell, 2005). On the other hand, if there is publication bias, the effect sizes and post hoc power values can be strongly overestimated (especially for small sample sizes). Unless the experiments have large sample sizes, which is not the case in Balcetis and Dunning (2010), the outcome of this analysis is not to be trusted.

Technique 2) may be justified in cases when one wants to estimate the mean of a distribution of heterogeneous true effect sizes, but it seems inappropriate to then treat that mean as a fixed effect for the publication bias test. If one believes that the effect sizes really are heterogeneous, then the publication bias test needs to further consider the uncertainty in the effect sizes. This is a more complicated analysis than was carried out by Francis (2012) and B&D's response. B&D's response may be correct that considering this type of heterogeneity of effect sizes is a good way to proceed, but none of us yet know exactly what that will entail or the result of the analysis.

B&D's response implies that if the probability of their experiment set is above the criterion of 0.1, then their findings are in the clear. With the two analyses they proposed, they end up with a probability for their findings of 0.116 and 0.163, respectively. The 0.1 criterion is somewhat arbitrary, but I agree that it should be respected if that is the basis for a classification. On the other hand, I would be reluctant to put much faith in a set of results that have only a 16% chance of occurring when the experiments are run properly. I would not, for example, use those findings to promote a controversial theory of visual perception.

In addition, it is important to realize that one or two experiments that successfully replicate the Balcetis and Dunning (2010) findings might push the set back below the criterion, even for their power analyses. For example, suppose that the pooled effect size remains unchanged but two new experiments reject the null hypothesis with power values of 0.6. Then, with technique 2) the probability of seven out of eight experiments rejecting the null hypothesis is .095. Counter intuitively; when a set of findings has a probability close to the criterion for publication bias, the desired experimental outcome (if one wants to believe in the effect) is a failure to replicate.

In general, it was proper for B&D's response to consider variations of the analysis because the test should be robust. In reality though, robustness is already built into the publication bias test. As Francis (2012) noted, there are good reasons to believe that experiments often overestimate effect sizes, so an exploration of robustness should consider that the true effect sizes might be smaller than what was reported. To avoid circularity, the publication bias test used in Francis (2012) gives the investigated experiments the benefit of the doubt by taking the reported effect sizes at face value. As a result, the test is quite conservative, even when a publication bias exists.

Areas of agreement

The findings in Balcetis and Dunning (2010) contained publication bias

B&D's response explained that there was an unpublished experiment with a null finding. Since not publishing an otherwise valid null finding is publication bias, the key result from the analysis in Francis (2012) is supported. As discussed below, this conclusion does hinge on the null finding being considered a valid study of wishful seeing.

The findings in Balcetis and Dunning (2010) are strengthened by sharing the originally unreported experiment

I agree with B&D's response that reporting their unpublished null finding improves the believability of their findings from a statistical point of view. It provides a better estimate of the effect size (even if the magnitude is relatively unchanged, the confidence interval is narrower), and it pulls the probability of their set of findings above the criterion used to indicate publication bias.

However, choosing to accept the previously unreported null finding as part of the set of experiments requires more information about the reasons it was not originally published. B&D's response stated that the reviewers and editor felt that the experiment had some flaws and requested a modified experiment, which is what was ultimately published.

If the reviewers and editor had legitimate criticisms regarding the unreported experiment, then it is inappropriate for B&D's response to include that study as part of their experiment set. If this is the case, then the experiment set reverts to what was analyzed by Francis (2012). Given the indication of bias in that set, we are left unsure about the validity of the main finding.

Determining whether the unpublished study in B&D's response is valid or invalid highlights one of the problems with publication bias: readers do not get to see the details of a suppressed study. Only when Balcetis and Dunning fully publish the properties of this experiment can subject matter experts determine whether it properly provides evidence for their substantive claims.

In contrast, suppose the editor and reviewers felt that the experimental design and methods were fine, but they had concerns about publishing a null finding. With this interpretation, B&D's response properly included the unpublished finding in the analysis, and they are correct that it brings the probability of the entire set to around 0.2, which is above the 0.1 criterion. Under this interpretation, the journal review process is a source of the publication bias in Balcetis and Dunning (2010).

Ultimately, authors are responsible for the content of their papers, but it is true that editors and reviewers can pressure authors to suppress valid but non-significant findings. I suspect

that Balcetis and Dunning (2010) made decisions similar to ones that many research psychologists make when faced with the task of getting an article published. I have every reason to believe that Balcetis and Dunning are honest researchers who diligently search for scientific truth to the best of their abilities. If Balcetis and Dunning are like other research psychologists I know (including myself from about a year ago), then they probably did not realize the impact of some of their choices. Very possibly the field needs wholesale changes in how psychological research is generated, analyzed, summarized, reviewed, and published. I would be delighted to work with Balcetis and Dunning to promote a better understanding of these important issues.

References

Balcetis E, Dunning D, 2010 "Wishful seeing: More desired objects are seen as closer" *Psychological Science* **21** 147-152

Balcetis E, Dunning D, 2012 "A false-positive error in search in selective reporting: A refutation of Francis" *i-Perception* **3** Author response

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

Kelley K, 2007 "Confidence intervals for standardized effect sizes: Theory, application, and implementation" *Journal of Statistical Software* **20**(8) 1-24

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