Response to the author's reply

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

25 March 2024

PNAS does not allow for a back-and-forth conversation, but I wanted to respond to the author's reply (Ongchoco, Water-Terrill, & Scholl, 2024—henceforth OWTS) to my letter (Francis, 2024) because I think they misunderstand the "excess success" analysis that was applied to their paper and because they give poor advice about interpreting replications.

1. OWTS note that my letter is one of several published analyses indicating publication bias (or other questionable practices) in published research. These numerous reports, unfortunately, reflect the general state of research in psychology, which has been experiencing a "replication crisis" for the past decade. My take is that many scientists, in the past, unwittingly used bad experimental designs and selectively reported results to support their conclusions. It is perhaps an indictment of the field that papers based on poor designs and selective reporting continue to be published, even with pre-registration, in high profile journals.
2. OWTS next make an "appeal to authority" argument by noting that the excess success analysis has been critiqued by leading experts in statistics and methodology. Such critiques do exist, but have been responded to by me (Francis, 2013b, 2016). There are also leading experts in statistics and methodology that largely support the excess success analysis (e.g., Ioannidis, 2013; Gelman, 2013). Scientists don't need to appeal to authorities, as they can look into the details for themselves (it is not that complicated). A good starting point is the special issue of the *Journal of Mathematical Psychology* where my full article (Francis, 2013a) described the details and properties of the excess success analysis, six experts provided commentaries about the article, and I wrote a response to their commentaries. My impression is that the discussion helped clarify some misunderstandings about the excess success analysis. However, OWTS misrepresent the views of some of those experts by presenting quotes out of context. For example, Vandekerckhove, Guan & Styrcula (2013) did say that the test was "all but useless", but they were talking about a special situation (where all published studies were subject to massive publication bias—a situation that I agree is essentially hopeless). Contrary to the representation given by OWTS, the start of the sentence containing that quote is, "While useful as a test for individual audits…", which is precisely the way I used the analysis in my *PNAS* letter.
3. Ignoring my response to those critiques, OWTS repeat some concerns about the analysis. For example, they argue that the excess success analysis is invalid because the findings I report suffer from publication bias itself. There might seem to be a satisfying sense of irony in this observation, but the argument simply is not true. There is bias in my reports of publication bias, but it is a kind of bias that does not

matter. I think it is intuitively obvious to empirical scientists that the impact of publication bias is related to the conclusions drawn from a set of studies. My unwillingness (or inability) to publish some non-significant experimental studies about visual afterimages does not alter the conclusions I might draw from published (significant) studies about short-term memory. This is a situation where there is publication bias that does not matter for the conclusions of the published studies. On the other hand, not publishing non-significant experimental results related to short-term memory might make my conclusions about short-term memory invalid. This is a situation where publication bias does matter. (Should I develop a unified theory of afterimages and short-term memory, the unpublished non-significant studies about afterimages might become relevant to my conclusions and would need to be published.) A bias that matters involves studies that are relevant to the conclusions. Clearly, the studies reported in Ongchoco, Walter-Terrill & Scholl (2023) are all relevant to their conclusions (why else would they report them?), so publication bias in that set (including studies that were not published) undermines their conclusions.

A second criticism raised by OWTS is that the false alarm rate of my analysis is unreasonably high. They note that applying the excess success analysis to 10 papers has a probability of 0.65 (calculated as $1-0.9^{10}$) of making at least one false alarm. Actually, the excess success test is rather conservative (Francis, 2013a) because it uses the reported statistics to estimate experimental power, so although I conclude "excess success" when the estimated replication probability is less than 0.1, the false alarm rate for any given set of studies is closer to 0.01 (so the probability of finding at least one false alarm across 10 applications of the test is around 0.1). Mathematical details aside, OWTS seem to mostly be criticizing the basic properties of statistical inference. There is always some risk of making a false alarm from noisy data; and this applies to their own experiments as well as to my analysis. The flip side of this aspect of inference is that a set of experiments that happens to produce results interpreted by the excess success analysis as indicating publication bias do so precisely because they look unusual. It *is* unusual to have experiments repeatedly produce *p*-values just a bit below, but never above, the 0.05 criterion. Perhaps Ongchoco *et al.* (2013) were just very unlucky and had almost every experiment barely produce a significant result; that *could* happen but I don't think scientists should have much faith in those results (precisely because they are so unusual as a set). This, too, is something that empirical scientists are already familiar with. Data that happen to generate a Type I error are unusual, but they look like real effects.

A third criticism by OWTS parrots an earlier misunderstanding (van Boxtel & Koch, 2016; Francis, 2016) about how the test is applied. I agree with OWTS that it would be "inappropriate…to compute post hoc power while averaging over tests exploring independent questions". I did not do that. It is appropriate to consider the probability that all the tests they used to support their conclusions would uniformly produce significant results. Their own data suggests that this should be a very uncommon outcome (probability of 0.011).

4. OWTS state that there was no publication bias in their set of studies. I hope they are wrong and that they misinterpreted some non-significant studies as being "pilot" studies rather than as evidence against their conclusions. The alternative (ignoring the very low probability of random sampling producing very unusual data) is that they did not follow their pre-registration plan, which many scientists would consider to be fraud.

    It is easy to misinterpret pilot studies. For example, suppose a scientist plans to gather data from n_1=n_2=200 participants for a two-sample t-test. As a methodological check the scientist gathers data from n_1=n_2=100 participants and runs a t-test. Data collection continues to the full sample size if the intermediate data looks promising, say, p<0.2 (not necessarily statistically significant) and in the right direction ($\bar{X}_1 > \bar{X}_2$); otherwise the experiment is halted and treated as a pilot study. If the null hypothesis is actually true in this case, then the proportion of completed studies that reject the null hypothesis (p<0.05) is about 0.17. [Simulation code to demonstrate this property is available at the OSF (Francis, 2023).] To the scientist using this approach across multiple experiments it might feel like they did nothing wrong; after all they reported all the completed experiments (no publication bias), and they collected exactly the planned sample sizes (no optional stopping or flexibility in data analysis). Nevertheless, the Type I error rate for the reported experiments is much inflated and the average effect size (d=0.12) is larger than reality (d=0). If the intermediate criterion is very stringent (e.g., p<0.001), then the Type I error rate across the completed studies is about 0.75, while the average effect size is d=0.25. Intuitively, this makes sense because the only studies run to completion are those that (just due to random sampling) have early data indicating big differences between the groups. These problems would get worse if there were additional intermediate checks during sampling.
5. OWTS misinterpret the implication of their within-article direct replication (Experiment 6 is a direct replication of Experiment 5, but with twice the sample size). Consider just the t-tests showing a significant anchoring effect for the No-Doorway condition. The t-statistic for Experiment 5 (t=2.03) suggests an estimated standardized effect size of d=0.203. For their Experiment 6, they used samples of size n_1=n_2=400, which for that effect size has power of 0.82. However, Experiment 6 found a weaker (but significant) effect (t=2.03, d=0.14). With this better estimate of the effect, we can reconsider the design of Experiment 5 to see that an experiment with n_1=n_2=200, only has power of 0.3, which is quite unsatisfactory. With that same effect size we conclude that the design of Experiment 6 only has power of 0.53. So, the probability that replication studies of the experiments (with the same sample sizes) will both produce significant outcomes is $0.3 \times 0.53 = 0.159$. So, rather than validating that the effects are reliable, their direct replication study highlights the implausibility of the experiments consistently producing significant outcomes.
6. In their final paragraph, OWTS suggest that actual (empirical) replication is the best test for replicability and that since their results replicated many times, they are confident that the results are reliable. I'm not going to have a debate with OWTS about whether calculus is correct; they are simply wrong. For the estimated effect

sizes in their data, and for the sample sizes they used, it is straightforward to apply calculus to the sampling distributions of null and alternative hypotheses to compute that the probability of seven experiments like the ones they report to all produce significant results is 0.011. If they believe the replication rate is much higher, then they must not trust their own data and instead believe that their experiments' effect sizes drastically underestimate reality. The burden of proof for such a claim is on them; and they have provided nothing to support that possibility.

7. Finally, their list of references is incorrect, as the name of a co-author is missing for an investigation of studies about object-based attention (Francis & Thunell, 2022). The correct citation is below.

References

1. Francis, G. (2013a). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, 57(5), 153-169.
2. Francis, G. (2013b). We should focus on the biases that matter: A reply to commentaries. *Journal of Mathematical Psychology*, 57, 190-195.
3. Francis, G. (2016). Confirming the appearance of excess success: Reply to van Boxtel and Koch (2016), *Psychonomic Bulletin & Review*, 23, 2010-2013.
4. Francis, G. (2023, December 20). Supplemental material for publication bias and anchoring. Retrieved from osf.io/g8r27
5. Francis, G. (2024). Evidence that publication bias contaminated studies of visual event boundaries and anchoring effects. *Proc Natl Acad Sci USA*., 121:13, e2320278121.
6. Francis, G. & Thunell, E. (2022). Excess success in articles on object-based attention. *Attention, Perception, & Psychophysics*, 84, 700-714.
7. Gelman, A. (2013). Interrogating p-values. *Journal of Mathematical Psychology*, 57(5), 188–189.
8. Ioannidis, J. P. A. (2013). Clarifications on the application and interpretation of the test for excess significance and its extensions. *Journal of Mathematical Psychology*, 57(5), 184–187.
9. Ongchoco, J. D. K., Water-Terrill, R. & Scholl, B. J. (2023) Visual event boundaries restrict anchoring effects in decision-making. *Proc Natl Acad Sci USA*., 120:44, e2303883120.
10. Ongchoco, J. D. K., Water-Terrill, R. & Scholl, B. J. (2024) Reply to Francis: Replicability, false alarms, and walking through doorways. *Proc Natl Acad Sci USA*., 121:13, e2401487121.
11. van Boxtel, J. J. A., & Koch, C. (2016). Reevaluating excess success in psychological science. *Psychonomic Bulletin & Review*, 23, 1602-1606.
12. Vandekerckhove, J., Guan, M., & Styrcula, S. A. (2013). The consistency test may be too weak to be useful: its systematic application would not improve effect size estimation in meta-analyses. *Journal of Mathematical Psychology*, 57(5), 170–173.