Reevaluating excess success in psychological science

Jeroen J. A. van Boxtel[1] · Christof Koch[2]

Published online: 12 February 2016

Abstract Francis (*Psychonomic Bulletin Review, 21*, 1180–1187, 2014) recently claimed that 82 % of articles with four or more experiments published in *Psychological Science* between 2009 and 2012 cannot be trusted. We critique Francis' analysis and point out the dependence of his approach on including the appropriate experiments and significance tests. We focus on one of the articles (van Boxtel & Koch, in *Psychological Science, 23*(4), 410–418, 2012) flagged by Francis and show that the inappropriate inclusion of experiments and tests have led Francis to mistakenly flag this article. We found that decisions about whether to include certain tests potentially affect 34 of the 44 articles analyzed by Francis. We further performed p-curve analyses on the articles discussed in Francis' analysis. We found that 9 of 44 studies showed significant evidential value, 11 studies showed insufficient evidential value, and 1 study showed evidence of p-hacking. Our reevaluation is important, because some researchers may have gained the false impression that none of the quoted articles in *Psychological Science* can be trusted (as stated by Francis). The analysis by Francis is most likely insufficient to warrant this conclusion for some articles and certainly is insufficient with respect to the study by van Boxtel and Koch (*Psychological Science, 23*, 410–418, 2012).

Keywords Statistics · Statistical inference

Introduction

In a recent article, Francis (2014) claimed that much of the research published in *Psychological Science* should not be trusted. We reevaluate Francis' study and identify the problems with his analysis in general and with respect to our study specifically (van Boxtel & Koch, 2012).

How does Francis support his claim? Francis (2014) uses a variation of the Test for Excess Significance (TES) of Ioannidis and Trikalinos (2007). In brief, he calculates the *post-hoc* power of each experiment reported in an article and multiplies these values to obtain the overall post-hoc power of that article. When this overall post-hoc power falls below 0.1, Francis flags the study as one whose results cannot be trusted.

A first point to emphasize is that the TES, as used by Francis, has garnered a lot of opponents. We will not reiterate many of the concerns, but it is worth emphasizing that statisticians criticize aspects ranging from the usefulness to the validity of the test (Morey, 2013; Simonsohn, 2013; but see Francis, 2013). These issues revolve around using the test to analyze the excess significance in single studies, but are not necessarily a problem for using the test in meta-analyses, as done by Ioannidis and Trikalinos (2007).

Putting these considerations aside, the TES used by Francis can only meaningfully be applied under certain conditions: (Requisite 1) To obtain enough power, Francis can only analyze papers with 4 or more experiments, because, assuming that average power is approximately 0.5, this would lead to an

Electronic supplementary material The online version of this article (doi:10.3758/s13423-016-1010-0) contains supplementary material, which is available to authorized users.

✉ Jeroen J. A. van Boxtel
j.j.a.vanboxtel@gmail.com; http://jeroenvanboxtel.com

Christof Koch
christofk@alleninstitute.org

[1] School of Psychological Sciences and Monash Institute of Cognitive and Clinical Neurosciences, Monash University, Clayton 3800, VIC, Australia

[2] Allen Institute for Brain Science, Seattle, WA, USA

overall power of $0.5^4 = 0.0625$, which is <0.1 (the cutoff used by Francis). (Requisite 2) The TES is only valid if the experiments included in the analysis all test the same hypothesis (Ioannidis, 2013). Furthermore, because Francis computes a single overall average effect size to estimate power for all experiments, the analysis used by Francis is only valid when the experiments are exact or close replications (Ioannidis & Trikalinos, 2007).

To calculate a post-hoc power for an *article*, Francis multiplies the post-hoc powers of the individual experiments. To calculate the post-hoc power for any one *experiment*, Francis multiplies the powers of the individual significance tests within that experiment. To derive a meaningful post-hoc power, one can only combine experiments, or significance tests within experiments, that test the same hypothesis (see Requisite 2). However, deciding which experiments (or significance tests) are testing the same hypothesis and therefore can be included in the analysis is not an easy task. It requires knowledge of the field (Francis, 2013; Johnson, 2013), or even of the mindset of the researcher when she conducted the experiments (Morey, 2013). We will discuss separately the choices that Francis made regarding the grouping of experiments (to calculate the overall power of the article) and the grouping of significance tests (to calculate the power of individual experiments) when he analyzed our article (van Boxtel & Koch, 2012).

Reanalysis of van Boxtel and Koch (2012)

In this comment, we will focus on one particular article (van Boxtel & Koch, 2012) that was flagged by Francis' analysis. We chose to focus on this article for several reasons. First, we wrote the article, and thus we are intimately familiar with the research topic and the analyses. We remain confident of its conclusions, in particular as they have been independently replicated (Vergeer, Boi, Öğmen, & Herzog, 2012). Second, the article requires a relatively complicated analysis in which Francis made several assumptions to meet the conditions allowing the use of the TES.

In the report (van Boxtel & Koch, 2012), we showed that observers perceive visual rivalry between two competing interpretations, even when there is no spatial overlap between the two sources of information: that is, visual rivalry without spatial conflict. This finding is important, because visual rivalry had previously always been found to rely on spatial conflict and was thus thought to rely on low-level, location-specific (spatial), visual mechanisms. Instead, this report suggests that a higher-level visual area could be the source of at least some forms of rivalry. In addition to the main experiment 1, several other experiments were conducted to further characterize this phenomenon (i.e., dependence on stimulus configuration and on object-based reference frames).

Choosing which experiments to group

In Francis' analysis, all four of the experiments that were conducted by van Boxtel and Koch (2012) were combined. According to Francis, the individual experiments have calculated powers of 0.52 (Experiment 1), 0.57 (Experiment 2), 0.65 (Experiment 3), and 0.36 (Experiment 4). Multiplying these values (when not rounded), leads to an overall post-hoc power of 0.071, implying that the probability of replicating our experiments with the same or greater success is 7.1 %, which Francis interprets as unlikely.

However, we argue that Francis cannot group all four experiments into one power analysis. For example, Experiment 4 is a clearly inappropriately included into the analysis. In this experiment, we tested whether the rivalry without spatial conflict that we found in experiment 1 was object-based or object-centered. We found no evidence for an object-centered effect and suggested that the rivalry is object-based. Francis took this experiment as a "replication" of our previous experiment. However, the results of Experiment 4 do not bear on the hypothesis of whether there exists rivalry without spatial conflict. Experiment 4 only qualified the type of rivalry as object-based, and not object-centered. Moreover, there was no condition in Experiment 4 that showed rivalry without spatial conflict. Therefore, Experiment 4 cannot be seen as a replication of experiment 1. When excluding this experiment, Francis' analysis becomes underpowered (by Requisite 1) and his conclusions unwarranted.

To summarize: (1) Francis incorrectly combines experiments with different methods, which test different hypotheses, and therefore fails to meet Requisite 2; (2) his analysis is underpowered had he performed the correct analysis, therefore failing to meet Requisite 1.

Choosing which significance tests to group

Similar to deciding which experiments to group into a power analysis, it is not always easy to select the right tests to include in the analysis of the power of individual experiments either (Francis, 2013; Johnson, 2013). We will focus on Experiment 1 from our article, because it included many tests, and selecting the correct tests is difficult. Again, Francis (2014) incorrectly combined various reported analyses in his calculation of the power of this experiment.

Experiment 1 tested whether rivalry occurred in conditions with different levels of object-based and retinotopic visual conflict. There were four conditions, each of which either had object-based visual conflict (O+) or not (O-) and retinal conflict (R+) or not (R-), resulting in four possible combinations (O + R+, O + R-, O-R+, O-R-).

The stimulus configuration was such that when observers did not perceive rivalry, they should have a perceptual bias towards a horizontal motion percept (coded as 0). When rivalry was perceived without any bias to one or the other percept, the perceptual biases would be 0.5. Therefore, to test whether rivalry was perceived, we tested all conditions compared with 0 and with 0.5, expecting rivalry to show perceptual biases significantly different from 0, but not significantly less than 0.5. To further check whether the conditions O + R+ and O + R- did not just show a large bias in the opposite direction, we also tested these conditions versus 1. The effect size and power for the tests are reported in Table 1. Importantly, the condition O + R- showed the pattern consistent with perceptual rivalry, indicating that rivalry without spatial conflict exists. O + R+ is the only other condition that appeared to show rivalry.

Through computer simulations, Francis calculated the overall power of this experiment to be 0.521. Note that, different from the TES as used by (Ioannidis & Trikalinos, 2007), Francis also includes tests that are predicted to be nonsignificant. When a test is significant and it was predicted to be significant, Francis considered this a "success," and similarly for tests that were nonsignificant when they were predicted to be nonsignificant. Therefore, in the last column in Table 1, we show the power of each test taking into account whether it was predicted to be significant according to Francis. We call this $\text{power}_{\text{success}}$. The $\text{power}_{\text{success}}$ is equal to the regular post-hoc power for tests that were predicted to be significant. However, to calculate the power of obtaining a non-significant result, we subtracted the power from 1. Therefore, when the expectation (according to Francis) is a nonsignificant result, $\text{power}_{\text{success =}}$ 1-power. By multiplying all values of $\text{power}_{\text{success}}$, we derive an overall power in a more direct way than Francis did. We find a $\text{power}_{\text{success}}$ of 0.527, which is very similar to Francis' power analysis (0.521). We note, however, that the average $\text{power}_{\text{success}}$ of all the tests is very high at 0.93, and based on a binomial test, finding a "success" in 8 of 8 tests is not significantly different from expected (p = 0.56). The calculated overall post-hoc power of the experiment therefore poorly reflects the strength of the findings.

Table 1 Overview of Cohen's d, post-hoc power, and post-hoc $\text{power}_{\text{success}}$ for the tests included in Francis' analysis

Test	d	power	Expect	$\text{power}_{\text{success}}$
O + R+ vs 0	6.930	1	+	1
O + R- vs 0	3.404	0.9999999	+	0.9999999
O-R- vs 0	0.665	0.3174227	-	0.6825773
O + R- vs 0.5	0.419	0.1558172	-	0.8441828
O-R+ vs 0.5	1.531	0.9182173	+	0.9182173
O-R- vs 0.5	2.141	0.9962284	+	0.9962284
O + R+ vs 1	2.487	0.9996460	+	0.9996460
O + R- vs 1	4.243	1	+	1
				Prod = 0.5269
				Mean = 0.9301

The column "Expect" shows a plus when a significant effect was expected and a minus when a nonsignificant effect was expected according to Francis

This poor reflection of the strength of the findings in Francis' analysis leads us to the major issue with this approach. Francis' analysis does not calculate the power of the experiment with respect to our hypothesis that there is rivalry without spatial conflict. Instead, his analysis tests a more elaborate conjoint hypothesis: O + R+ and O + R- should be significantly different from zero, and the O-R+ and O-R- should not, and O + R+ and O + R- should not be significantly different from 0.5, whereas O-R+ and O-R- should. In other words, this more complicated hypothesis tests whether there is rivalry in O + R+ and O + R-, and additionally whether there is no rivalry in O-R+ and O-R-. It therefore tests four separate hypotheses, of which only one was of main interest to us. Therefore, this approach violates Requisite 1. Importantly, by incorrectly including tests that are immaterial to the hypothesis under scrutiny, the power the experiment is severely underestimated.

What would be the best way to assess the power of experiment 1, as related to the hypothesis about rivalry without spatial conflict? The only tests that matter are those that assess whether there is rivalry in the condition with object-based conflict but no retinotopic conflict (i.e., condition O + R-). The condition without any type of conflict (O-R-) should serve as a baseline compared with which O + R- should be significantly increased. A paired t test reveals this to be the case (t(6) = 4.37, p = 0.005, d = 1.6514, power = 0.95; in the article we reported a more conservative comparison against O-R+, reaching the same conclusion). To ensure the perceptual bias is not too extreme, one can test that the perceptual bias is smaller than 1 (power ≈ 1, Table 1). Multiplying the power of these tests gives a post-hoc power of 0.95, which is much higher than the 0.52 reported by Francis.

In Experiment 3, Francis also combines tests that investigate different hypotheses. In this experiment, only the group-motion condition was a replication of experiment 1. In this condition, a clear baseline is lacking, which is why we can compute the perceptual bias only relative to 0 and not to another baseline condition. Compared with 0, the perceptual bias is significantly different (t(11) = 5.15, p < 0.0005, Cohen's d = 1.49, $\text{power}_{\text{success}}$ = 0.997). The perceptual bias is not different from 0.5 (t(11) = 1.6, p > 0.13, $\text{power}_{\text{success}}$ = 0.69). Multiplying these values yields an overall power of 0.69, again higher than the power reported by Francis (0.65).

Conclusions after the reanalysis

We conclude that there is no reason to doubt that there is rivalry without spatial conflict. The studies in our article were sufficiently powered, and the TES—when conducted more specifically to investigate our hypothesis—does not reveal a bias. The TES as performed by Francis erroneously included Experiment 4. When Experiment 4 is excluded from the analysis, the post-hoc power over the first 3 experiments is increased to 0.19. Addressing the unnecessary inclusion of several tests in Experiment 1 increases the post-hoc power of that experiment to 0.95. Combined, this reanalysis lifts the power over the first 3 experiments to 0.35. Given this analysis there is no indication that our report is biased.

Sources of bias

When a study is flagged by the TES, it suggests that a bias is present in a report (or group of reports). However, it does not determine the origin of the bias. In the next section, we will discuss different potential sources of bias (i.e., publication bias, harking, and p-hacking) in relation to our published article and, later, to the other articles discussed by Francis (2014).

Publication bias

A publication bias is introduced by publishing more of the found significant than nonsignificant findings. This applies to published articles as a whole, but also to individual experiments within an article: Authors can introduce a publication bias by not reporting experiments that were not significant (although one may consider this p-hacking, see below).

Murayama, Pekrun, and Fiedler (2014), however, showed that when several successful replications are reported within one article. This can generally be taken as evidence of an effect, even when not all experiments are reported, although the reported effect size might be inflated. This is, after all, the basic premise at the heart of science, *i.e.,* repeated confirmation by independent means of some hypothesis.

Harking

HARKing stands for *H*ypothesizing *A*fter the *R*esults are *K*nown. This is an issue when one does not have a hypothesis when the experiment is conducted, but one reports the results later as if there was. We cannot prove that we did not perform HARKing, but because we based our experimental protocol and hypothesis on a well-established technique advanced by Herzog and colleagues (Boi, Ogmen, Krummenacher, Otto, & Herzog, 2009), there would seem little evidence for it. Further support for our findings comes in the form of a conference report with very similar findings (Vergeer et al., 2012).

One interesting aspect to discuss is Experiment 4, where we did not have a hypothesis, and the results we report and the interpretation we gave may be seen as harking. However, we see this as an exploratory experiment. Although this was not explicitly mentioned in the article, it is obvious from the introduction of this experiment that it was not a hypothesis-driven study and should be considered exploratory. As mentioned before, the conclusions of this experiment do not bear on the main finding of the paper, namely that there exists rivalry without spatial conflict.

Had we been clearer in our description of this experiment, perhaps Francis would not have considered this experiment as part of our set of predictions. He would then not have included it in the TES, and our report would not have been flagged. An important lesson is thus to identify experiments (and tests) as exploratory when they are.

P-hacking

P-hacking is trying multiple statistical analyses until obtaining the desired results (generally accompanied by only reporting those results). This behavior will lead to a disproportionate amount of reported p-values just below the significant threshold (generally 0.05) (Simonsohn, Nelson, & Simmons, 2014). This disproportionality can be subjected to a significance test, which could indicate evidence of p-hacking.

We can never prove that we did not p-hack, but to provide some support we ran the analysis explained by Simonsohn et al. (2014). We calculated the p-curve based on the data reported in our article. The p-curve analysis showed no evidence of p-hacking in our data, and in fact shows strong evidence for evidential value (Table S1).

What about the other articles in psychological science?

In relation to the TES, we have argued that one of the major difficulties is correctly grouping experiments (and significance tests) when calculating the post-hoc power of the article. Obviously, all articles are potentially affected by the choice of experiments to group, but which articles could also be affected by the choice of significance tests within one experiment? We looked at the 44 studies discussed by Francis. Any study in which Francis used multiple tests to construct the post-hoc power of an experiment could be affected by this issue. Furthermore, any study where both ANOVA results and multiple *t* tests were reported covering the same data also are potentially affected. All of these studies are marked by 1 in column 2 in Table S1. A total of 34 of the 44 studies are potentially affected.

It is important to realize that we did not test whether another choice of grouping would affect the results—such a choice would have required knowledge of the field—but one can see that most studies are potentially affected by the choice of tests. Potentially Francis' analysis may contain false positives (and negatives) just because of the choice of tests he included.

Are any of the other issues relevant to the other papers discussed by Francis (2014)? Harking is not easy to detect, nor is selective reporting of experiments. P-hacking can be investigated with the p-curve (Simonsohn et al., 2014). Therefore, we analyzed all 44 papers discussed by Francis (2014) with the p-curve and report the outcomes in Table S1. Our evaluation revealed that only one article was highlighted as exhibiting evidence of p-hacking, although many studies were highlighted as having inadequate evidential value as calculated according to Simonsohn et al. (2014).

Conclusions

Francis cannot conclude that our article (van Boxtel & Koch, 2012) contains excess success, because his test is underpowered for the number of replications reported in our study. He also misrepresents (i.e., underestimates) the power of our experiments by including several tests that are immaterial to our conclusions. Similar arguments may hold for other studies in the analysis reported by Francis (2014) (see Table S1, column 2).

Even though we criticize Francis' analysis, we do not want to claim that report biases in the field of psychology do not exist. They do, as they do in many other fields. But we want to highlight that the analyses performed by Francis (Francis, 2014) do not warrant this conclusion for at least some individual reports like ours. Other researchers (e.g., Morey, 2013) claim that Francis' analyses do not warrant this conclusion for any report.

Acknowledgments The authors thank Uri Simonsohn, Richard Morey, and one anonymous reviewer for their insightful comments on the manuscript.

References

Boi, M., Ogmen, H., Krummenacher, J., Otto, T. U., & Herzog, M. H. (2009). A (fascinating) litmus test for human retino- vs. non-retinotopic processing. *Journal of Vision, 9*(13), 5. doi:10.1167/9.13.5. 1–11.

Francis, G. (2013). We should focus on the biases that matter: A reply to commentaries. *Journal of Mathematical Psychology, 57,* 190–195.

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin Review, 21*(5), 1180–1187. doi:10.3758/s13423-014-0601-x

Ioannidis, J. P. (2013). Clarifications on the application and interpretation of the test for excess significance and its extensions. *Journal of Mathematical Psychology, 57,* 184–187.

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials (London, England), 4*(3), 245–253. doi:10.1177/1740774507079441

Johnson, V. E. (2013). On biases in assessing replicability, statistical consistency and publication bias. *Journal of Mathematical Psychology, 57,* 177–179.

Morey, R. D. (2013). The consistency test does not–and cannot–deliver what is advertised: A comment on Francis (2013). *Journal of Mathematical Psychology, 57,* 180–183.

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review, 18*(2), 107–118. doi:10.1177/1088868313496330

Simonsohn, U. (2013). It really just does not follow: Comments on Francis (2013). *Journal of Mathematical Psychology, 57,* 174–176.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547. doi:10.1037/a0033242

van Boxtel, J. J., & Koch, C. (2012). Visual rivalry without spatial conflict. *Psychological Science, 23*(4), 410–418. doi:10.1177/0956797611424165

Vergeer, M., Boi, M., Öğmen, H., & Herzog, M. H. (2012). *Binocular suppression occurs in object-centered coordinates.* Paper presented at the VSS.