***Test for Excess Significance.*** Another power-based test for publication bias is the Test for Excess Significance (Ioannidis & Trikalinos, 2007). This test estimates the number of expected studies with statistical significance given some anticipated effect size (usually the naïve meta-analytic estimate), then compares that expectation against the number of observed significant results. A significant test suggests censoring of nonsignificant results or the manipulation of results into statistical significance.

This test has a number of weaknesses that prevent its inclusion in the main text. It has poor statistical power (Ioannidis & Trikalinos, 2007), and the validity of its *p*-value rests on strong, perhaps unwarranted assumptions about researcher behavior (Morey, 2013). Like other tests for bias, its results may be spurious when there exists genuine between-study heterogeneity. The reader is urged to interpret these results with considerable caution.

The results are summarized in Table S1. TES indicates a significance of excess significance in the full sample of experiments of aggressive behavior and its best-practices subsample. This test is also significant in the full sample of experiments of aggressive affect and near the significance threshold in the best-practices subsample.

***Multiplicative-error meta-regression models.*** Random-effects meta-analysis models heterogeneity across studies through the use of an additive error term τ2. An alternative strategy for handling heterogeneity in meta-regression is instead to fit models with a multiplicative error term φ. (see, e.g., Moreno et al., 2009).

We fit additional multiplicative-error models. The results are summarized in Table S2. The most significant change is that the PET estimate for best-practices aggressive behavior reaches statistical significance in these models, and the PEESE estimate increases to *r* = .17.

**Sensitivity analyses.** Some data points appeared to be outliers, having unusually high effect sizes and thus considerable influence on the meta-analytic adjustments. We report these here.

In experiments of aggressive affect, one best-practices study (Ballard & Wiest, 1996) appeared to be an outlier, having a very large effect size and low precision. Exclusion of this outlier brought the estimators into greater agreement, reducing the naïve estimates (*r* = .27, fixed- and random-effects, *I2* = 0.01, [0.00, 62.8]), the *p*-uniform estimate (*r* = .20), and the *p*-curve estimate (*r* = .19), while increasing the PET (*r* = -.01, *I2* = 0.00, [NA, NA]) and PEESE (*r* = .17, *I2* = 0.00, [NA, NA]) estimates. We think that the exclusion of this outlier is therefore a good idea, as it prevents PET from returning a negative effect size. Again, these null-set confidence intervals on *I2* indicate unusual homogeneity of residuals after adjusting for small-study effects.

[Other influential observations include …]

Table S1. Results of the Test for Excess Significance.

Table S2. Naïve, PET, and PEESE estimates in fixed-effect dispersion models.

References

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

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* (5), 180 - 183. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022249613000291