March 2nd, 2015

Comparing *p*-curve Results Aggregating *pp*-values with Fisher's vs Stouffer's method

Uri Simonsohn
University of Pennsylvania
uws@wharton.upenn.edu

Leif D. Nelson
University of California, Berkeley

Joseph P. Simmons
University of Pennsylvania

Abstract:
Starting on March 2015, *P*-curve's online app computes the statistical significance of *p*-curves relying on Stouffer's rather than Fisher's method for aggregating *p*-values. Here we report the results from our original paper (Simonsohn, Nelson, & Simmons, 2014) under both methods side-by-side, noting the results are highly comparable. We then contrast the sensitivity of *p*-curve results to incorporating a few extremely significant but fake results originally published in retracted work by Larry Sanna. Stouffer's method is much less sensitive to those extreme results.

Skipping background information

We write this document assuming readers are familiar with *p*-curve in general and Simonsohn et al.'s (2014) paper in particular. We do not repeat the description of concepts explained in that article, e.g., *pp*-values, nor provide any details on the source of the data or simulations re-analyzed here. We focus on reporting the original Fisher's Method based results, and the new ones based on Stouffer's method.

The JPSP Demonstration (Figure 3)

In Simonsohn et al. (2014), we compiled test results from studies expected to contain evidential value, and studies expected to lack evidential value. All studies came from articles published in the *Journal of Personality and Social Psychology*. Below we re-print Figure 3 in our article, Figure 1 here. The statistical results reported in it rely on Fisher's method. In Figure 2 here we reprint the output from the *p*-curve's App 3.0, which relies on Stouffer's method.

For the set of studies expected to contain evidential value, Fisher's method leads to an overall test for right-skew of $\chi^2(44)=94.2$, $p<.0001$, while Stouffer's method to Z=-3.5, $p=.0002$. The 33% power null is not rejected with either method ($p=.51$ and $p=.73$ respectively).

For the set of studies expected to *lack* evidential value, Fisher's method leads to an overall test for left-skew that barely rejects the null, with $\chi^2(40)=58.2$, $p=.031$. Stouffer's method rejects the null more strongly, Z=-2.64, $p=.0042$.

Figure 1. *Reprint of Figure 3 in Simonsohn et al (2014).*

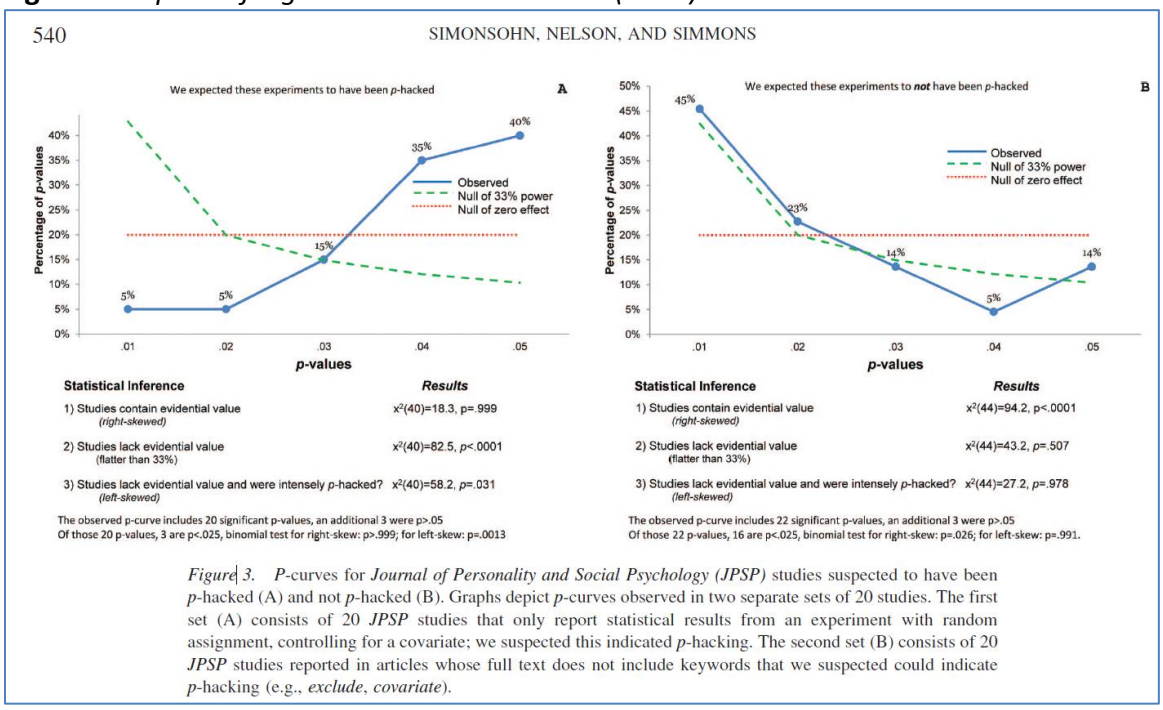
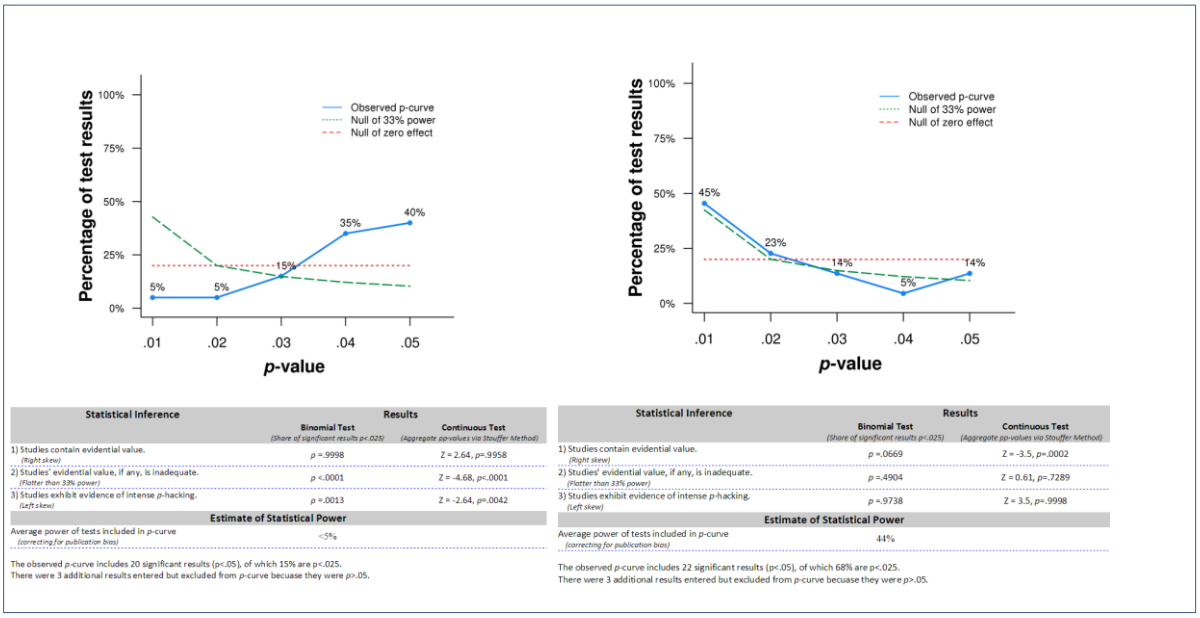


Figure 3. *P*-curves for *Journal of Personality and Social Psychology (JPSP)* studies suspected to have been *p*-hacked (A) and not *p*-hacked (B). Graphs depict *p*-curves observed in two separate sets of 20 studies. The first set (A) consists of 20 *JPSP* studies that only report statistical results from an experiment with random assignment, controlling for a covariate; we suspected this indicated *p*-hacking. The second set (B) consists of 20 *JPSP* studies reported in articles whose full text does not include keywords that we suspected could indicate *p*-hacking (e.g., *exclude*, *covariate*).

Figure 2. *Results from p-curve App 3.0, for same tests.*



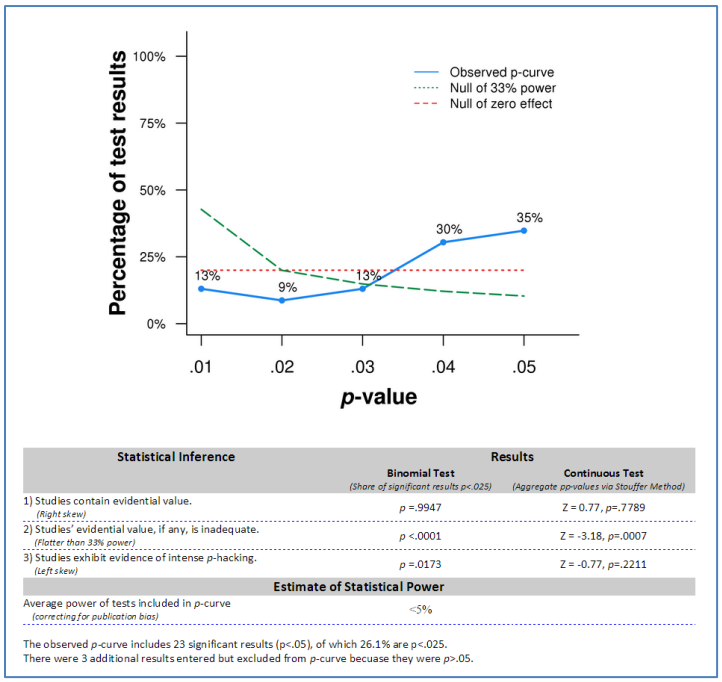
Adding fake studies to left-skewed *p*-curve

To demonstrate how Stouffer's method is less influenced by extreme observations, and why this may be desirable, we added to the 20 studies resulting in a left skew (first panel of Figures 1 and 2), results from three lab experiments known to involve fabricated data (Sanna, Chang, Miceli, & Lundberg, 2011). While fake-data need not lead to impossibly small *p*-values, impossibly small *p*-values will often arise in fake-data.

In Sanna's case, the three critical *p*-values for the fake studies were: .011. .00001, and .000004. If we add these three low *p*-values to the *p*-curve containing the 20 results suspected to lack evidential value, the overall shape remains left-skewed, and Stouffer's method continues to conclude the curve is very far from significantly right skewed (p=.78). Fisher's method, in contrast, leads to an overall $\chi^2(46) = 57.3$ that is dramatically lower than before, p=.12.

To be clear, we do not use Stouffer's method to be robust to fake data per-se, but rather, to extreme observations that can be obtained through many sources. These include fraud, human error (in the original reporting or *p*-value selection), confounds in original design, etc. *P*-curves seeks to establish if a set of studies, *generally* contains evidential value. A technique that is less sensitive to few extreme studies seems better aligned with the motives behind carrying out *p*-curve analyses in the first place.

Figure 3. *P-Curve combining left-skew JPSP curve with three fake studies by Sanna et al.*



Statistical Inference	Results	
	Binomial Test (Share of significant results p<.025)	Continuous Test (Aggregate pp-values via Stouffer Method)
1) Studies contain evidential value. (Right skew)	*p* =.9947	Z = 0.77, *p*=.7789
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	*p* <.0001	Z = -3.18, *p*=.0007
3) Studies exhibit evidence of intense *p*-hacking. (Left skew)	*p* =.0173	Z = -0.77, *p*=.2211
Estimate of Statistical Power		
Average power of tests included in *p*-curve (correcting for publication bias)	<5%	

The observed *p*-curve includes 23 significant results (p<.05), of which 26.1% are p<.025.
There were 3 additional results entered but excluded from *p*-curve becuase they were *p*>.05.

Type 1 and Type 2 Errors with *p*-curve with Stouffer and Fisher's method

Simonsoh et al (2014), in Figure 6 of that paper, reported the power *p*-curve had to detect evidential value, and the false-positive rate concluding that *p*-curve was flatter than the 33% power *p*-curve, for *p*-curves including 5, 10 and 20 studies, with the underlying power of those studies being 33%, 50% and 80%. Figures 4 and 5 below contrasts those results when *p*-curve is analyzed with Fisher vs. Stouffer's method.

The figures show that Stouffers exhibits a minimal loss of power (<2 percentage points), and a similarly small lower false-positive rate.

Figure 5. *Power to detect Right-Skew using Fisher vs. Stouffer method for aggregating pp-values*

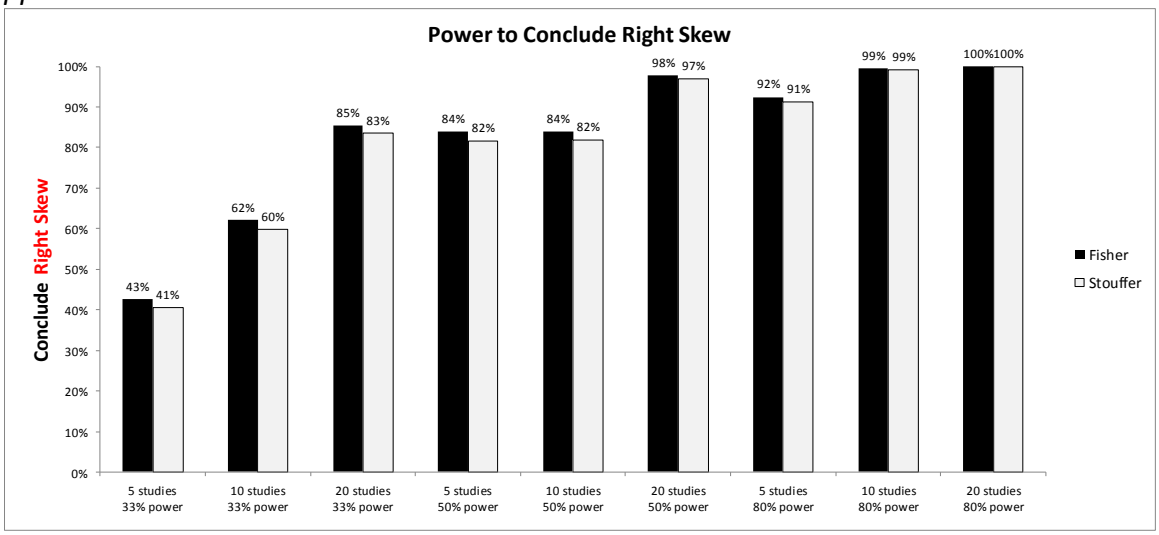
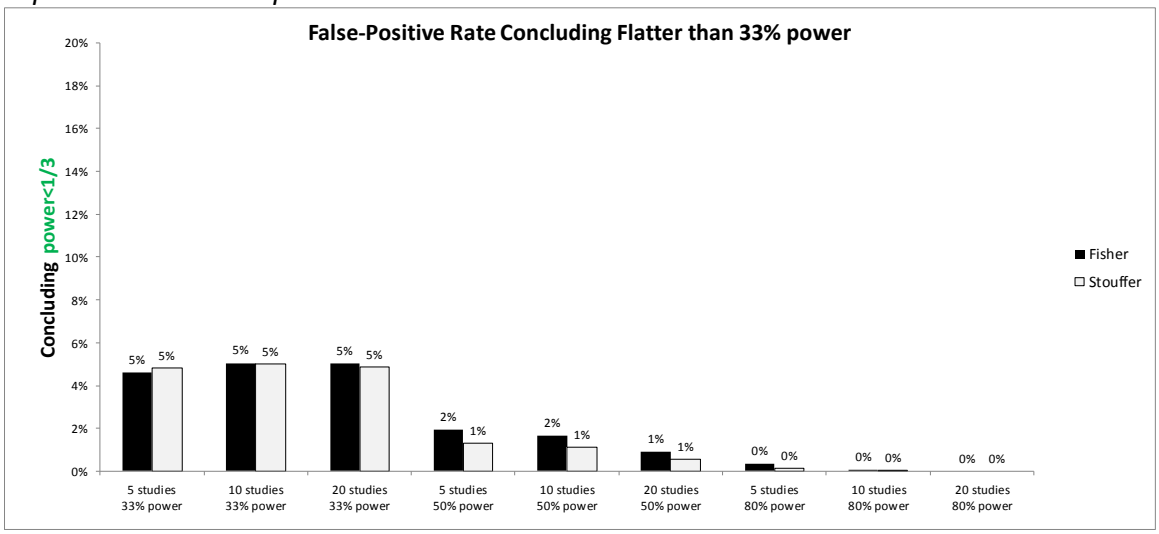


Figure 6. *False-Positive Rate concluding the observed p-curve is flatter than that expected under 33% power.*



References

Sanna, L. J., Chang, E. C., Miceli, P. M., & Lundberg, K. B. (2011). "Retracted: Rising up to Higher Virtues: Experiencing Elevated Physical Height Uplifts Prosocial Actions". *Journal of Experimental Social Psychology, 47*(2), 472-476.

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.