P-curves are better at effect size estimation than trim-and-fill (and Michael Jordan is better at free throws than I am)

Daniel Lakens[1]

[1]Affiliation not available

April 17, 2023

the WINNOWER

MATHEMATICS

P-curves are better at effect size estimation than trim-and-fill (and Michael Jordan is better at free throws than I am)

DANIEL LAKENS

CORRESPONDENCE:
lakens@gmail.com

DATE RECEIVED:
June 24, 2015

DOI:
10.15200/winn.143515.56858

ARCHIVED:
June 24, 2015

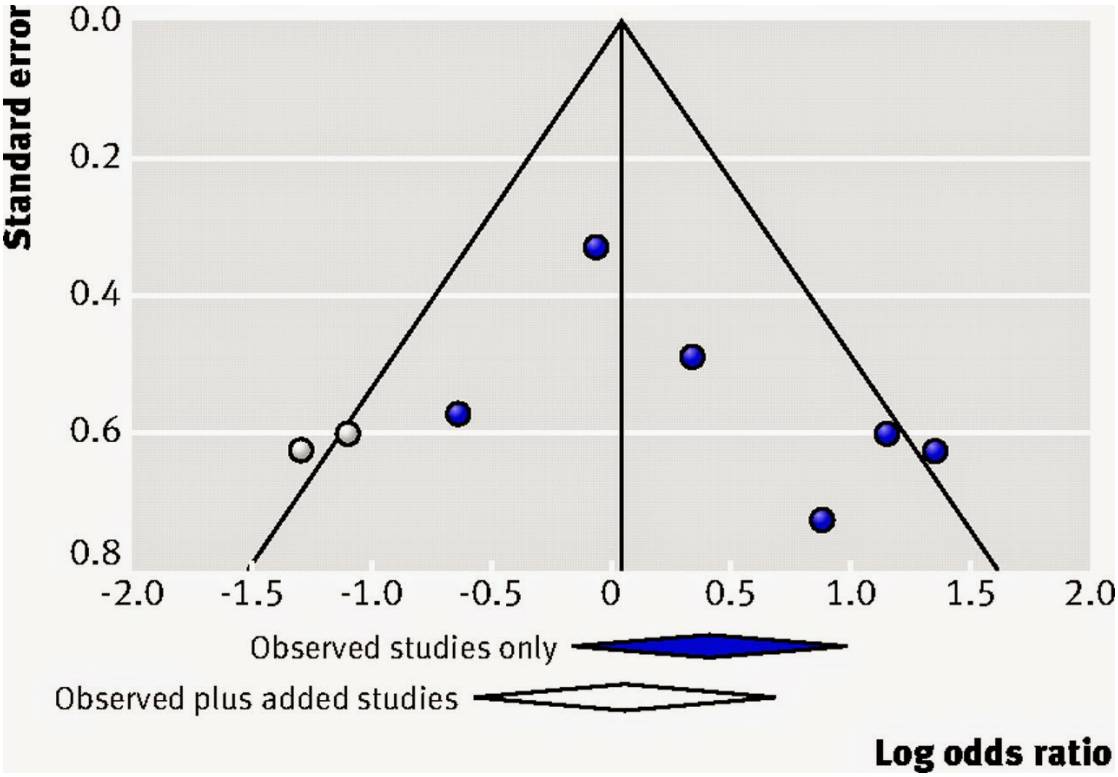
KEYWORDS:
p-curve, meta-analysis, trim-and-fill, PET-PEESE

CITATION:
Daniel Lakens, P-curves are better at effect size estimation than trim-and-fill (and Michael Jordan is better at free throws than I am), *The Winnower* 2:e143515.56858 , 2015 , DOI: 10.15200/winn.143515.56858

I like p-curve analyses. They are a great tool to evaluate whether sets of studies have evidential value or not. In a recent paper, Simonsohn, Nelson, & Simmons (2014) show how p-curve analyses can correct for publication bias using only significant results. That's pretty cool. However, trim-and-fill is notoriously bad at unbiased effect size estimation. I think what we really need is a comparison between state-of-the-art methods to correct effect size estimates for publication bias. I dip a toe in the water at the end of this post, but can't wait until someone does a formal comparison between p-curve and techniques like p-uniform (a technique developed independently by Marcel van Assen and colleagues) and meta-regression techniques like PET-PEESE analyses by Stanley and Doucouliagos.

If you wanted to know how good a basketball player Michael Jordan is, you wouldn't compare how much better he is at free throws than I am. The fact that *p*-curve outperforms trim-an-fill methods in a simulation where studies are not published purely based on their *p*-value is only interesting for historic reasons (trim-and-fill is well-known, even though it is now considered out-dated), but not interesting when you want to compare state-of-the-art techniques to correct meta-analytic effect sizes for publication bias.

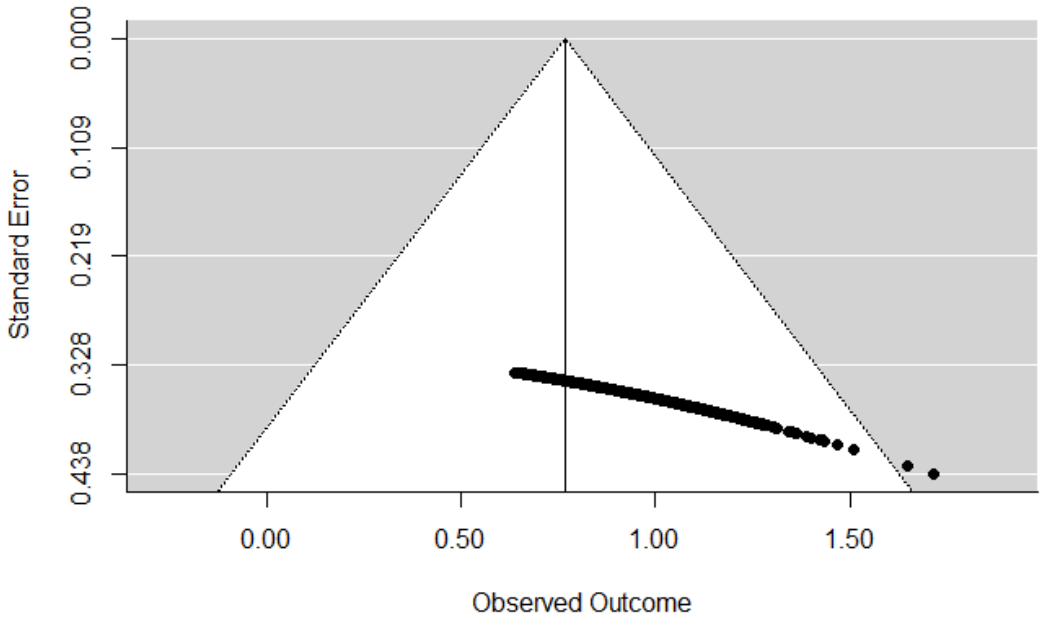
Trim-and-fill is not created to examine publication bias caused by effects that yield non-significant *p*-values. Instead, it's created for publication bias caused by effects that are strongly (perhaps even significantly!) in the opposite direction of the remaining studies or the effect a researcher wants to find. Think of a meta-analysis on the benefits of some treatment, where the 8 studies that revealed the treatment actually makes people feel much worse are hidden in the file drawer. In the picture below (source) we can see how trim-and-fill assumes there were two missing studies in a meta-analysis (of which one was significant in the opposite direction, because the diagonal lines correspond to 95% confidence intervals).



How does trim-and-fill work? The procedure starts by removing ('trimming') small studies that bias the meta-analytic effect size, then estimates the true effect size, and ends with 'filling' in a funnel plot with studies that are assumed to be missing due to publication bias.

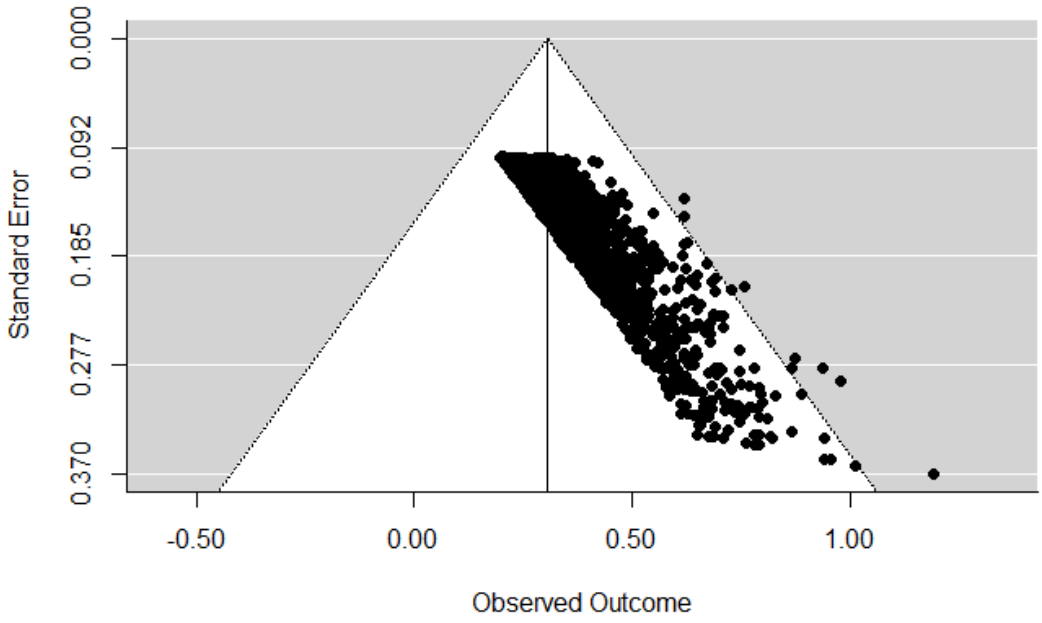
However, it is already known that trim-and-fill is not very good under many realistic publication bias scenarios (such as when publication bias is caused based on the height of the *p*-value of the effect, which is arguable the most common source of publication bias in psychology). The method is criticized for its reliance on the strong assumption of symmetry in the funnel plot, and when publication bias is induced by a *p*-value boundary (such as in the simulation by Simonsohn et al), the trim-and-fill method does not perform well enough to yield a corrected meta-analytic effect size estimate that is close to the true effect size (Peters, Sutton, Jones, Abrams, & Rushton, 2007; Terrin, Schmid, Lau, & Olkin, 2003). *When the assumptions are met*, it can be used as a sensitivity analysis (with little difference between the corrected and uncorrected effect size estimate indicating publication bias is unlikely to change the conclusions of the meta-analysis), but even then, if there are differences between the corrected and uncorrected effect size estimate, researchers should not report the corrected effect size estimate as an estimate of the true effect size (Peters et al., 2007).

I've adapted the simulation by Simonsohn et al (2014) to look at only one meta-analysis at a time, and re-run one analysis underlying Figure 2a. It allows me to visualize a single simulation, and below is a nice demonstration of why it is important to visualize your data. We see the simulated data is quite different from what you typically see in a meta-analysis. Note that the true effect size in this simulation is 0, but power is so low, and publication bias is so extreme, the meta-analytic effect size estimate from a normal meta-analysis is very high (d = 0.77 instead of the true d = 0). It's excellent *p*-curve analysis can accurately estimate effect sizes in such a situation, even though I don't hope anyone will ever perform a meta-analysis on studies that look like the figure below. As I said, I totally believe Michael Jordan is going to win at free-throws from me (especially if he can stand half as far from the basketball ring).



Let's increase the variability a little by adding more studies of different sizes (ranging from N=40 to N=400). The great R script by Simonsohn et al makes this very easy. We see two things in the picture below. First of all, re-running the simulation, while allowing for sample sizes of up to 400 reduces the effect size overestimation. Although it is no-where near 0, this demonstrates how running only hugely underpowered studies can immensely increase publication bias. It also demonstrates that running larger studies will in itself not fix science - we also need to publically share all well-designed studies to prevent publication bias.

The second thing we see is that it start to look a little bit (albeit a very little bit) more like a normal distribution of effect sizes from mainly underpowered studies that suffer from publication bias (yes, that's a very specific definition of 'normal', I know). Where trim-and-fill believed 0 studies were missing in the simulation above, it thinks a decent number of studies (around 20%) are missing below (not nearly enough) and it even adjusts the meta-analytic effect size estimate a little (although again not nearly enough, which is to be expected, given that it's assumptions are violated in this simulation).



How would *p*-curve compare to more state-of-the-art meta-analytic techniques that correct for publication bias? My favorite (with the limited knowledge I have of this topic at the moment) is PET-PEESE meta-regression. For an excellent introduction through application, including R code, see this paper by Carter & McCullough.

I ran a PET (precision-effect-test) in both situations above (see R script below). PET really didn't know what to do in the first example above, probably because regression based approaches need more variability in the standard error (so the presence of larger and smaller studies). However, in the test on the second simulation, PET performed very well, estimating the true meta-analytic effect size corrected for publication bias did not significantly differ from 0. So, under the right conditions, both *p*-curve and PET-PEESE meta-regression can point out publication bias and provide an unbiased effect size estimate of 0, when the true effect size is indeed 0.

So how would Michael Jordan do compared to Magic Johnson? And how would *p*-curve do against *p*-uniform or PET-PEESE meta-regression? I'd love to see a comparison, especially focusing on which technique will work best under which assumptions. For example, does *p*-curve work better when there are only small studies, but does PET-PEESE do better when publication bias is less extreme? Publication bias is the biggest problem a quantitative science can have, and it's clear that with so many improved techniques to calculate unbiased effect sizes estimates, we have important tools to draw better inferences from the scientific literature. Let's see how to combine these tools to make the best inferences under which circumstances. If there's a rogue statistician out there who wants to help us researchers select the best tool for the job, a detailed comparison of these different recent approaches to correct meta-analytic effect size estimates for publication bias would be an extremely useful project for those cold winter days ahead.

Carter, E. C., & McCullough, M. E. (2014). Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.00823

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, *26*(25), 4544-4562.

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

Stanley, T. D & Doucouliagos, Hristos. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*, 60-78. doi:10.1002/jrsm.1095

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*(13), 2113-2126.