

The p -curve is Not “Just Fine”

Blake McShane, Ulf Bockenholt, and Karsten Hansen

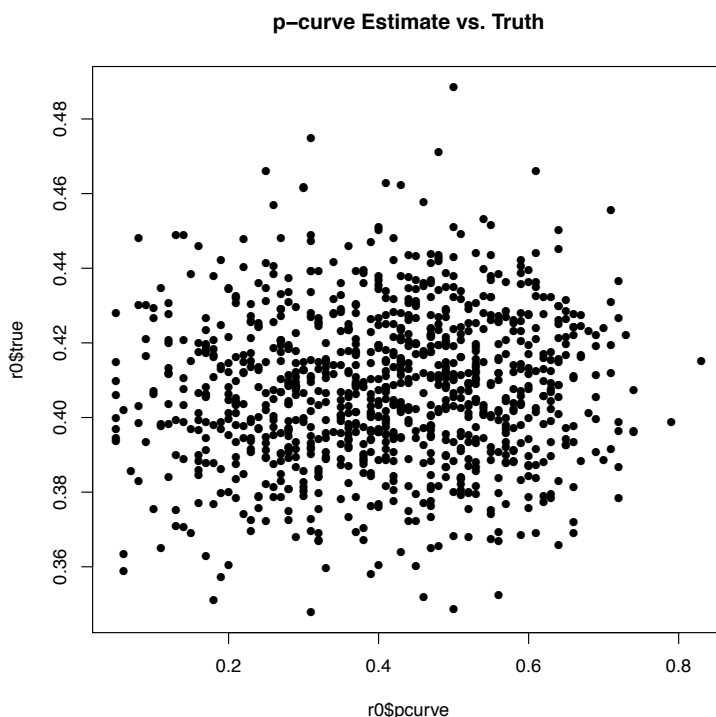


Figure 1: “P-curve Handles Heterogeneity Just Fine” (<http://datacolada.org/67>) according to Simonsohn, Nelson, and Simmons but their own simulation makes clear there is nothing “fine” about p -curve estimates: perfect estimates would fall on the 45-degree line while p -curve estimates look like random scatter.

Note: To reproduce the above figure, run the R code (<http://datacolada.org/appendix/Colada%2067%20-%20Shorter%20-%202017%2012%2019.R>) associated with the SNS blog post and then run: `plot(r0$pcurve, r0$strue, pch=16, main="p-curve Estimate vs. Truth")`.

Preface: This is a long form response to a recent post of Andrew Gelman. The original idea for the post was to feature a dialogue between the three of us and Uri Simonsohn with additional commentary by Andrew; the below reflects that. Unfortunately, the dialogue did not work out. Andrew proceeded with his own post and solicited brief comments from each of us. We note that several of us on opposing teams are friends and admire one another’s work, and nothing in this comment has a personal element to it.

In “ p -Curve and Effect Size” (2014), SNS purport to introduce a new method, the so-called p -curve, that adjusts for the upward bias in meta-analytic effect size estimates that results from publication bias (i.e., the tendency for statistically significant and directionally consistent results to be over-represented in the published literature). However, as we have written (McShane, Bockenholt, & Hansen, 2016), the method is not new and it is terribly flawed in the sense that it

cannot be relied upon to produce reasonable—yet alone the claimed definitive—adjustments in practice.

Nonetheless, SNS have continued to defend the use of the p -curve (<http://datacolada.org/61>)—even when effect sizes are heterogeneous (<http://datacolada.org/67>)—and have also argued against the typicality of heterogeneity in psychological research (<http://datacolada.org/63>). Since last June, we have engaged in a long and detailed correspondence with Uri (Simonsohn) over these issues both prior and subsequent to the publication of these blog posts. SNS view our differences as mere disagreement about “how to effectively communicate methods issues to a general audience.” We believe they run deeper than that, involving three important *matters of fact*:

[1] *Mathematical*: The p -curve employs a model for the observed data that is identical to that of Hedges (1984) and yields upwardly-biased and highly inaccurate estimates of the population average effect size (PAES) when effect sizes are heterogeneous. Thus, the p -curve cannot be relied upon in practice to provide valid or definitive adjustments for publication bias.

[2] *Methodological*: Meta-analytic research has for decades focused on estimating the PAES and estimating it accurately as measured by mean square error (MSE) or similar quantities. However, when it suits, SNS *selectively report* various novel estimands in lieu of the PAES as well as various novel model evaluation metrics in lieu of accuracy.

[3] *Empirical*: A host of recent data convincingly demonstrates that effect size heterogeneity is the norm in psychological research. SNS argue otherwise in a manner that misses the point and that confuses observed vs. unobserved and systematic vs. unsystematic heterogeneity.

[1] **Mathematical**:

As we wrote, SNS reinvented the wheel: the p -curve employs a model for the observed data that is identical to that of Hedges (1984)—a seminal paper which initiated the long and continuing literature on selection methods reviewed by, for example, Hedges and Vevea (2005) both of which are cited and dismissed by SNS.

Specifically, both Hedges (1984) and the p -curve assume observed t -statistics follow a truncated non-central t -distribution with common non-centrality parameter and where the truncation point is determined by a one-sided statistical significance threshold. In the language of selection methods, the data model is a non-central t -distribution with common non-centrality parameter and the selection model is only statistically significant and directionally consistent studies are published.

However, Hedges (1984) estimates the single model parameter (i.e., the effect size) using the principled maximum likelihood approach whereas SNS employ an *ad hoc* improvised procedure based on the Kolmogorov-Smirnov statistic. Consequently, due to the properties of maximum likelihood estimators, the p -curve estimate will naturally be inferior to the Hedges (1984) estimate (except perhaps in very small samples).

Further, when effect sizes are heterogeneous, both Hedges (1984) and the p -curve yield upwardly-biased and highly inaccurate estimates of the PAES—the central quantity of interest in meta-analytic research. This upward bias is a simple consequence of Jensen’s Inequality.

These straightforward mathematical facts are covered in detail in our paper.

They, in tandem with the fact that moderate to large effect size heterogeneity is the norm in psychological research (see [3]), means the p -curve cannot be relied upon in practice to provide valid or definitive adjustments for publication bias. Instead, we advocate using more sophisticated selection methods for adjustment, but solely for sensitivity analysis rather than to obtain a single definitive adjustment (see our paper for details).

[2] Methodological:

When evaluating a model, one cannot choose estimands and evaluation metrics to suit one’s purposes: some estimands and evaluation metrics are superior to others, and *selective reporting* is as much an issue in methodological research as it is in empirical research. However, in defending the p -curve, SNS engage in *selective reporting* or what might be eponymously termed methodological p -hacking in deference to their own coinage.

In particular, meta-analytic research has for decades focused on estimating the PAES and estimating it accurately as measured by mean square error (MSE) or similar quantities. However, when it suits, SNS *selectively report* various novel estimands in lieu of the PAES as well as various novel model evaluation metrics in lieu of accuracy. Any evaluation that fails to consider how accurate estimates of the central estimand of interest is at best woefully incomplete and at worst a distortion.

For example, in “ p -Curve and Effect Size”, SNS make two claims. First, they claim via simulation studies that the p -curve provides unbiased effect size estimates when effect sizes are homogenous; evidence for this claim is provided most strongly in their Figure 4. However, rather than plotting the mean estimate across their various simulation iterations, they plot the median estimate. This is of course not necessarily *selective reporting* but is certainly *simply wrong*: bias by definition refers to the comparison of truth and mean not truth and median. When this figure is redrawn but using the mean rather than median as is proper, it is clear the p -curve is biased.

Second, they claim “ p -curve is robust to heterogeneity in effect size across studies (see Supplement 2 for more details and additional variations).” As discussed above, due to Jensen’s Inequality it most certainly is not. In their Supplement 2, SNS *selectively report* a different estimand, moving the goal post from the PAES to an alternative and novel effect size measure and again focusing solely on bias in the context of a simulation study.

As another example, consider the recent SNS post (<http://datacolada.org/67>). In that post:

(i) SNS move the goal post yet again and *selectively report* a novel estimand in lieu of the PAES, namely the simple average of the power of the historic studies included in the meta-analysis

(they mention this quantity in passing at the end of “*p*-Curve and Effect Size” but this is clearly not the focus of that paper as indicated by the title and figures contained therein).

(ii) SNS again *selectively report* bias, and only bias, as the evaluation metric rather than accuracy.

(iii) As the scatterplot at the beginning of this document—which comes directly from the simulation study designed by SNS and featured in their blog post and which even uses their own novel estimand—makes clear, the *p*-curve produces extremely inaccurate estimates: perfect estimates would fall on the 45-degree line while *p*-curve estimates look like random scatter.

(iv) Uri shared this post with us before it went live and we in turn shared this scatterplot. We noted that showing an estimator is unbiased or has low bias—as SNS do for the settings considered in their post—is neither necessary nor sufficient for showing that the estimator is good (i.e., accurate) as reflected by the standard relationship $MSE = Bias^2 + Variance$. In other words, focusing solely on the bias term yields a misleading and incomplete view of the accuracy / performance of an estimator.

We can draw an analogy to psychological research. Consider a researcher running a lab experiment who collected two dependent variables, analyzed them, found one “worked” (i.e., yielded the expected finding) and the other didn’t, and then wrote up a discussion mentioning only the first DV. This researcher would—rightly and objectively—not just be incompletely or inaccurately reporting the results but instead would be distorting them, and this would be especially the case were the two DVs measures of the same / a similar construct rather than an orthogonal one. Analogously, the (i) the SNS post shows that when the *p*-curve is used to estimate average power the bias term is small/negligible in the settings they considered, (ii) our scatterplot shows that the sum of the two terms is large (thus in particular the second term is large because the first is small) in those settings, (iii) these two related constructs that in composite form MSE / accuracy ought to be evaluated jointly, (iv) to focus solely on the small/negligible bias after being made aware of the large variance is not just incompletely or inaccurately reporting results but distorting them.

Nonetheless, SNS chose to run with the title “*P*-curve Handles Heterogeneity Just Fine” and did not acknowledge the scatterplot or accuracy issue in the post.

In an earlier post (<http://datacolada.org/61>) which sparked our long correspondence, we also noted to Uri that he was *selectively reporting* yet another evaluation metric, moving the goal post from accuracy to Type I error. In that post, SNS demonstrate the singularly unilluminating fact that when one particular selection method (the so-called three-parameter selection method) is misspecified, it performs poorly with respect to Type I error; this is not illuminating for misspecified models in general perform poorly (our response to that blog post can be found here: <http://incurablynuanced.blogspot.com/2017/06/meta-analysis.html>).

SNS are of course free to propose novel estimands and novel evaluation metrics. For various reasons, we do not find their proposed ones particularly compelling. For example, we find average power uninteresting because (i) it is intrinsically bound up with the null hypothesis

significance testing paradigm which we find generally implausible in the social sciences (McShane *et al.*, 2017), (ii) it is conditional on the sample sizes of the historical studies and says nothing about future studies, and (iii) average power and the effect size are, under homogenous effect size models like the p -curve, simple transformation of one another and as such are redundant. Similarly, we find bias uninteresting when considered alone because we are interested in accuracy and we find Type I error uninteresting for reason (i) of the prior sentence.

Regardless, what they (or anyone for that matter) ought not do is engage in model evaluation that distorts results as in the post (<http://datacolada.org/67>). They also ought not omit evaluation of the central estimand of interest in meta-analysis (i.e., the PAES) as well as evaluation with regards to measures of estimation accuracy (and perhaps also measures of uncertainty calibration, for example of estimated standard errors and confidence intervals). Again, any evaluation that fails to consider how accurate estimates of the central estimand of interest is at best woefully incomplete and at worst a distortion.

[3] Empirical

A host of recent data convincingly demonstrates that effect size heterogeneity is the norm in psychological research. In particular, it is rife and large in comprehensive meta-analyses of psychological studies of the sort published in *Psychological Bulletin*, the premier outlet for meta-analyses in psychology. For example, van Erp *et al.* (2017) examined heterogeneity estimates from 705 meta-analyses published there between 1990 and 2013 and found a median I^2 of 71% (I^2 is the proportion of the observed variance due to variance in true effect sizes rather than sampling error). Similarly, Stanley *et al.* (2017) surveyed 200 recent meta-analyses published there and found a median I^2 of 74%.

More interesting and more important, heterogeneity persists—and to a reasonable degree—even in large-scale replication projects such as the Many Labs project (Klein *et al.*, 2014) and Registered Replication Reports (RRRs; Simons, *et al.* 2014) where rigid, vetted protocols with identical study materials are followed across labs in a deliberate attempt to eliminate it. Despite these extreme efforts to achieve homogeneity, a reasonable degree of heterogeneity (I^2 on the order of 40%) has been found in such projects (Klein *et al.*, 2014, McShane, Böckenholt, & Hansen, 2016; Hagger *et al.*, 2016; Eerland *et al.*, 2016).

From a substantive perspective, this is astounding! It is also astounding from another purely statistical perspective. As is well known, when heterogeneity is in fact nonzero but there are a relatively small number of studies / labs (as is unfortunately the case in Many Labs and RRRs), standard heterogeneity estimates are biased downwards and, in particular, estimates of zero heterogeneity result implausibly often (see, for example, Chung, Rabe-Hesketh, Gelman, Liu, and Dorie, 2013). For this and related reasons, the Type I error of the significance test of zero heterogeneity is often inflated above the nominal $\alpha = 5\%$ level and power is low (Huedo-Medina *et al.*, 2006; Ioannidis *et al.*, 2007). Thus, even if one believes homogeneity plausible and one believes in significance testing, one should be very skeptical of this particular significance test unless the number of studies is indeed very large.

In sum, there are two key points here. First, large-scale meta-analyses like those published in *Psychological Bulletin* feature a large degree of heterogeneity. Second, in large scale replication efforts such as Many Labs and RRRs, we should—for substantive reasons (i.e., extreme measures employed to minimize heterogeneity) and statistical reasons (i.e., estimates and significance tests of heterogeneity perform poorly in a manner that falsely suggests the absence of heterogeneity)—expect to see little heterogeneity in such data; the very fact we do observe a reasonable degree of heterogeneity is compelling evidence heterogeneity cannot be avoided in psychological research—even when every effort is taken to eliminate or minimize it!

Importantly, this justifies our recommendation that the p -curve cannot be relied upon to provide valid or definitive adjustments for publication bias in meta-analysis due to its poor performance when effect sizes are heterogeneous. We further note these adjustments are most important and useful precisely in the settings where these procedures perform worst—large-scale meta-analyses—and such adjustments are entirely unnecessary in the Many Labs and RRRs settings where publication bias is absent.

Uri downplays heterogeneity in a recent blog post (<http://datacolada.org/63>) but misses the point: regardless of the degree of heterogeneity found in large-scale replication projects, that there is any whatsoever in a setting where every effort is taken to eliminate it is both substantively interesting and strong evidence that it simply cannot be eliminated; this is especially the case given the statistical difficulties in detecting heterogeneity when there is a relatively small number of studies / labs. Indeed, to argue for his point that homogeneity is the norm in psychological research, it is not sufficient to argue or provide evidence that there is low or even no heterogeneity where one would *not* expect to find it. Instead, one must show evidence that there is low or no heterogeneity precisely where one *would* expect to find it. We also note that Uri *selectively reports* one particular effect (anchoring) in the post rather than examining all thirteen effects studied in the data and believe it important to reiterate all the statistical issues with the classical heterogeneity estimates and significance tests employed by Uri in this post.

[3] Redux

In their blog post (<http://datacolada.org/67>), SNS attempt by analogy to explain why we find it important to emphasize that the mathematical fact that the p -curve produces an upwardly biased effect size estimates when effect sizes are heterogeneous. In doing so, they confuse observed/unobserved and systematic/unsystematic heterogeneity. They write:

Imagine that an effect is much stronger for American than for Ukrainian participants. For simplicity, let's say that all the Ukrainian studies are non-significant and thus excluded from p-curve, and that all the American studies are $p < .05$ and thus included in p-curve. P-curve would recover the true average effect of the American studies. Those arguing that p-curve is biased are saying that it should recover the average effect of both the Ukrainian and American studies, even though no Ukrainian study was included in the analysis

This is not at all the parameter we would advocate recovering. Obviously, were we to know there was a systematic difference in two populations as here, we would advocate altering the model to accommodate this difference (e.g., adding a binary American vs. Ukrainian variable as a

predictor). We also note that, even if we did not possess such American vs. Ukrainian information but we believed there might be, for example, two distinct effect sizes in the population, we would advocate altering the model to accommodate latent classes. We also emphasize that, more realistically, there would not be one American effect size and one Ukraine effect size but rather a distribution of effect sizes for each country and we would advocate altering the model still further to accommodate this (again, whether or not we possessed the putative American / Ukrainian information).

In sum, it is unobserved unsystematic heterogeneity that the above evidence makes clear is the norm in psychological research and that is discussed in our paper, this post, and their posts (<http://datacolada.org/63> and <http://datacolada.org/67>). However, the American/Ukrainian example they use in their latter post to describe heterogeneity is an illustration of observed systematic heterogeneity which is obviously of a fundamentally different sort.

Conclusion

To conclude, we have delineated three issues related to the p -curve in publications and blog posts written by SNS. We believe the vast majority of what we have written above is entirely lacking in controversy. Thus, it is natural to ask why we have made the effort to write this post and to correspond for over half a year with Uri over these rather jejune issues?

The answer is simple: as statisticians who work each and every day with psychologists, we care about clearing up the record when it is incomplete or in error.

In the event it is not clear, it is important to understand that SNS are regarded as statistical / methodological gurus in the psychology community. Further, their blog serves as a highly influential platform for the dissemination of their usually good but sometimes misguided ideas.

Collectively, we serve on the editorial boards of half a dozen psychology journals and we have seen SNS blog posts treated as gospel by authors, editors, and reviewers in the review process. This is far from an ideal state of affairs. Their blog is not peer-reviewed and it does not allow “post-publication peer review” via a comments section. Further, although SNS officially have a policy of soliciting feedback on posts by authors they cite and allowing such authors to post responses, our experience with this has been decidedly unsatisfactory.

We believe psychology deserves better. As SNS has shown with regard to empirical work, openness and transparency is incredibly important, and there are practices about which there are matters of right and wrong. The same is true with regards to methodological work.

References:

- Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics*, 9, 61–85.
- Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 145–174). Chichester, England: John Wiley & Sons.
- Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F., and Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I² index? *Psychological methods*, 11(2):193.
- Ioannidis, J. P., Patsopoulos, N. A., and Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *BMJ: British Medical Journal*, 335(7626):914.
- Klein, R. A., Ratliff, K., Nosek, B. A., Vianello, M., Pilati, R., Devos, T., Galliani, E. M., Brandt, M., van 't Veer, A., Rutchick, A. M., Schmidt, K., Bahnik, S., Vranka, M., IJzerman, H., Hasselman, F., Joy-Gaba, J., Chandler, J. J., Vaughn, L. A., Brumbaugh, C., van swol, L., Wichman, A., Packard, G., Brooks, B., Cemalcilar, Z., Storbeck, J., Bocian, K., Levitan, C., Bernstein, M. J., Krueger, L. E., Eisner, M., Davis, W. E., Nier, J. A., Nelson, A. J., Steiner, T. G., Mallett, R., Thompson, D., Huntsinger, J. R., Morris, W., Skorinko, J., and Kappes, H. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology* 45, 3, 142–152.
- McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11(5), 730-749.
- McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2017). Abandon statistical significance. Tech. rep., Northwestern University.
- Simons, D. J., Holcombe, A. O., and Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science* 9, 5, 552–555.
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9, 666–681
- Stanley, T. D., Carter, E. C., Doucouliagos, H. (2017). What Meta-Analyses Reveal about the Replicability of Psychological Research. Deakin Laboratory for the Meta-Analysis of Research, Working Paper, November 2017.

van Erp, S., Verhagen, A. J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990-2013. *Journal of Open Psychology Data*, in press.

Blog posts:

<http://datacolada.org/61>

<http://incurablynuanced.blogspot.com/2017/06/meta-analysis.html>

<http://datacolada.org/63>

<http://datacolada.org/67>