Datacolada Post [61] Why *p*-curve excludes *p*s>.05
Response of Blakeley B. McShane, Ulf Böckenholt, and Karsten T. Hansen

Summary

We offer a two-page (six point) response to the recent blogpost by Simonsohn, Simmons, Nelson (SSN) on adjusting for publication bias in meta-analysis. We disagree with many of the points raised in the blogpost for reasons discussed in our recent paper on this topic [MBH2016]. Consequently, our response focuses on clarifying and expounding upon points discussed in our paper and provides a more nuanced perspective on selection methods such as the three-parameter selection model (3PSM) and the *p*-curve (a one-parameter selection model (1PSM)).

We emphasize that all statistical models make assumptions, that many of these are likely to be wrong in practice, and that some of these may strongly impact the results. This is especially the case for selection methods and other meta-analytic adjustment techniques. Given this, it is a good idea to examine how results vary depending on the assumptions made (i.e., sensitivity analysis) and we encourage researchers to do precisely this by exploring a variety of approaches. We also note that it is generally good practice to use models that perform relatively well when their assumptions are violated. The 3PSM performs reasonably well in some respects when its assumptions are violated while the *p*-curve does not perform so well. Nonetheless, we do not view the 3PSM or any other model as a panacea capable of providing a definitive adjustment for publication bias and so we reiterate our view that selection methods—and indeed any adjustment techniques—should at best be used only for sensitivity analysis.

Datacolada Post [61] Why *p*-curve excludes *p*s>.05
Response of Blakeley B. McShane, Ulf Böckenholt, and Karsten T. Hansen

Note: In the below, "statistically significant" means "statistically significant and directionally consistent" as in the Simonsohn, Simmons, Nelson (SSN) blogpost. In addition, the "*p*-curve" refers to the methodology discussed in SNS2014 that yields a meta-analytic effect size estimate that attempts to adjust for publication bias.[1]

Point 1: It is impossible to definitively adjust for publication bias in meta-analysis
As stated in MBH2016, we do not view the three-parameter selection model (3PSM) or any other model as a panacea capable of providing a definitive adjustment for publication bias. Indeed, all meta-analytic adjustment techniques—whether selection methods such as the 3PSM and the *p*-curve or other tools such as trim-and-fill and PET-PEESE—make optimistic and rather rigid assumptions; further, the adjusted estimates are highly contingent on these assumptions. Thus, these techniques should at best be used only for sensitivity analysis.
[For more details in MBH2016, see the last sentence of the abstract; last paragraph of the introduction; point 7 in Table 1; and most especially the entire Discussion.]

Point 2: Methods discussions must be grounded in the underlying statistical model
All statistical models make assumptions. Many of these are likely to be wrong in practice and some of these may strongly impact the results. This is especially the case for selection methods and other meta-analytic adjustment techniques. Therefore, grounding methods discussions in the underlying statistical model is incredibly important for clarity of both thought and communication.
SSN argue against the 3PSM assumption that, for example, a p=0.051 and p=0.190 study are equally likely to be published; we agree this is probably false in practice. The question, then, is what is the impact of this assumption and can it be relaxed? Answer: it is easily relaxed, especially with a large number of studies.
We believe the *p*-curve assumptions that (i) effect sizes are homogenous, (ii) non-statistically significant studies are entirely uninformative (and are thus discarded), and (iii) a p=0.049 study and a p=0.001 study are equally likely to be published are also doubtful. Further, we know via Jensen's Inequality that the homogeneity assumption can have substantial ramifications when it is false—as it is in practically all psychology research.
[For more details in MBH2016, see the Selection Methods and Modeling Considerations sections for grounding a discussion in a statistical model and the Simulation Evaluation section for the performance of the p*-curve.]*

Point 3: Model evaluation should focus on estimation (ideally across a variety of settings and metrics)
SSN's simulation focuses solely on Type I error—a rather uninteresting quantity given that the null hypothesis of zero effect for all people in all times and in all places is generally implausible in psychology research (occasional exceptions like ESP notwithstanding). Indeed, we generally expect effects to be small and variable across people, times, and places. Thus, "$p < 0.05$ means true" dichotomous reasoning is overly simplistic and contributes to current difficulties in replication. Instead, we endorse a more holistic assessment of model performance—one that proceeds across a variety of settings and metrics and that focuses on estimation of effect sizes and the uncertainty in them. Such an evaluation reveals that the 3PSM actually performs quite well in some respects—even in SSN's Cases 2-5 and variants thereof in which it is grossly misspecified (i.e., when its assumptions are violated; see Point 6 below).
[For more details in MBH2016, see the Simulation Design and Evaluation Metrics subsection.]

Point 4: The statistical model underlying the* p*-curve is identical to the model of Hedges, 1984 [H1984]
Both the *p*-curve and H1984 are one-parameter selection models (1PSM) that make identical statistical assumptions: effect sizes are homogenous across studies and only studies with results that are statistically significant are "published" (i.e., included in the meta-analysis). Stated another way, the statistical model underlying the two approaches is 100% identical and hence if you accept the assumptions of the *p*-curve you therefore accept the assumptions of H1984 and vice versa.
The only difference between the two methods is how the single effect size parameter is estimated from the data: H1984 uses principled maximum likelihood estimation (MLE) while *p*-curve minimizes the Kolmogorov-

[1] The same authors have developed a distinct methodology also labelled *p*-curve that attempts to detect questionable research practices. This note does not comment on that methodology.

Smirnov (KS) test statistic. As MLE possesses a number of mathematical optimality properties; easily generalizes to more complicated models such as the 3PSM (as well as others even more complicated); and yields likelihood values, standard errors, and confidence intervals, it falls on SSN to mathematically justify why they view the proposed KS approach to be superior to MLE for psychology data[2].
[For more details in MBH2016, see the Early Selection Methods and p*-methods subsections.]*

Point 5: Simulations require empirical and mathematical grounding
For a simulation to be worthwhile (i.e., in the sense of leading to generalizable insight), the values of the simulation parameters chosen (e.g., effect sizes, sample sizes, number of studies, etc.) and the data-generating process must reflect reality reasonably well. Further still, there should ideally be mathematical justification of the results. Indeed, with sufficient mathematical justification a simulation is entirely unnecessary and can be used merely to illustrate results graphically.
The simulations in MBH2016 provide ample mathematical justification for the results based on: (i) the optimal efficiency properties of the maximum likelihood estimator (MLE; Simulation 1), (ii) the loss of efficiency resulting from discarding data (Simulation 2), and (iii) the bias which results from incorrectly assuming homogeneity as a consequence of Jensen's Inequality (Simulation 3). We remain uncertain about the extent to which Cases 2-5 of the SSN simulations reflect reality and thus seek mathematical justification for the generalizability of the results. Nonetheless, they seem of value if viewed solely for the purpose of assessing the 3PSM model estimates when that model is misspecified.
[For more details in MBH2016, see the Simulation Evaluation section.]

Point 6: The 3PSM actually performs quite well in SSN's simulation—even when misspecified.
Only in Case 1 of the SSN simulation is the 3PSM properly specified (and even this is not quite true as the 3PSM allows for heterogeneity but the simulation assumes homogeneity). SSN show that when the 3PSM is misspecified (Cases 2-5), its Type I error is far above the nominal α=0.05 level. We provide further results in the figures that follow.

- The blue bars in the left panel of Figure 1 below reproduce the SSN result. We also add results for the 1PSM as estimated via KS (*p*-curve) and MLE (H1984). As can be seen, the Type I error of the 1PSM MLE remains calibrated at the nominal level. In the right panel, we plot estimation accuracy as measured by RMSE (i.e, the typical deviation of the estimated value from the true value). As can be seen, the 3PSM is vastly superior to the two 1PSM implementations in some cases and approximately equivalent to them in the remaining ones.
- In Figure 2, we change the effect size from zero to small (d=0.2); the 3PSM has much higher power and better estimation accuracy as compared to the two 1PSM implementations.
- In Figure 3, we return to zero effect size but add heterogeneity (τ=0.2). The 1PSM has uncalibrated Type I error for all cases while the 3PSM remains calibrated in Case 1; in terms of estimation accuracy, the 3PSM is vastly superior to the two 1PSM implementations in some cases and approximately equivalent to them in the remaining ones[3].
- In Figure 4, we change the effect size from zero to small and add heterogeneity. The 3PSM generally has similar power and better estimation accuracy as compared to the two 1PSM implementations (indeed, only in Case 1 does the 1PSM have better power but this comes at the expense of highly inaccurate estimates).

In sum, the 3PSM actually performs quite well compared to the two 1PSM implementations—particularly when the focus is on estimation accuracy as is proper; this is especially encouraging given that the 1PSM is correctly specified in all five cases of Figures 1-2 while the 3PSM is only correctly specified in Case 1 of the figures. Although these results favor the 3PSM relative to the two 1PSM implementations, we reiterate our view that selection methods—and indeed any adjustment techniques—should at best be used only for sensitivity analysis.

---

[2] Both MLE and KS are asymptotically consistent and thus asymptotically equivalent for the statistical model specified here. Consequently, any justification will likely hinge on small sample properties which can be mathematically intractable for this class of models. Justifications based on robustness to model specification are not germane here because if a different specification deemed more appropriate, the model would be re-specified according to this more appropriate specification and that model estimated.
[3] A careful reading of SNS2014 reveals that the *p*-curve is not meant to estimate the population average effect size. As shown here and in MBH2016, it cannot as no 1PSM can. This is important because we believe that the heterogeneous effect sizes (i.e., $\tau > 0$) are the norm in psychology research.

Figure 1: d=0 & tau=0

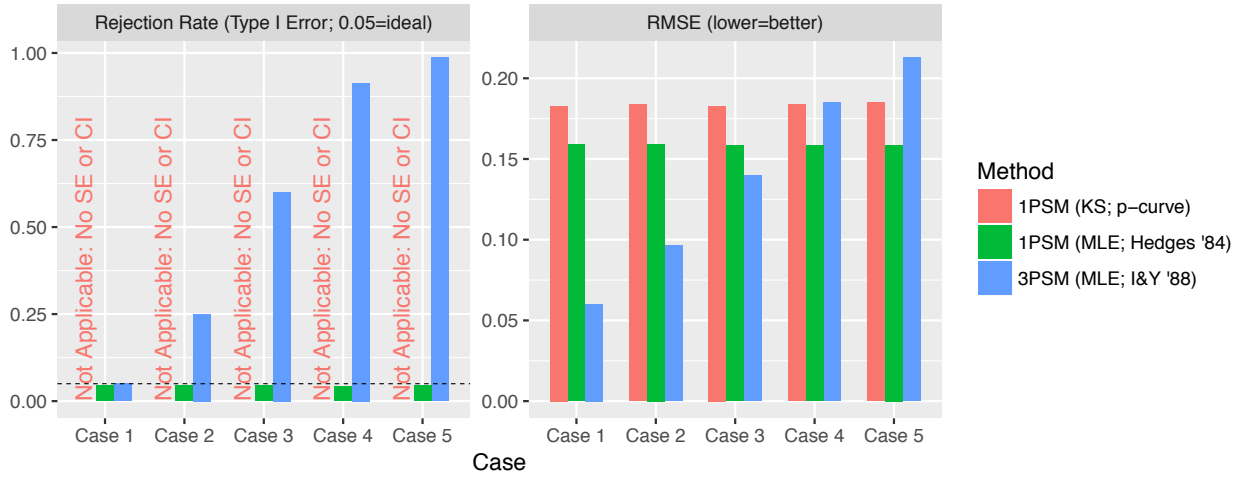

Figure 1 Zero Homogeneous Effect Size. The blue bars in the left panel replicate the SSN blogpost results showing the 3PSM has Type I error far above the nominal α=0.05 level when it is misspecified as in Cases 2-5. On the other hand, the Type I error of the 1PSM MLE remains calibrated at the nominal level.

In the right panel, we plot estimation accuracy as measured by RMSE (i.e, the typical deviation of the estimated value from the true value). As can be seen, the 3PSM is vastly superior to the two 1PSM implementations in some cases and approximately equivalent to them in the remaining ones.

In the left panels of this and the below figures, we note the *p*-curve does not natively produce SEs or CIs. While they can be obtained via the bootstrap, this is computationally infeasible for a simulation study where, for example, 1000 bootstrap iterations would be required for each of the 5000 simulation iterations. Moreover, the bootstrap poses other problems for the p-curve in this setting [*For more details in MBH2016, see the last paragraph of the* p*-Methods subsection.]*.

Figure 2: d=0.2 & tau=0

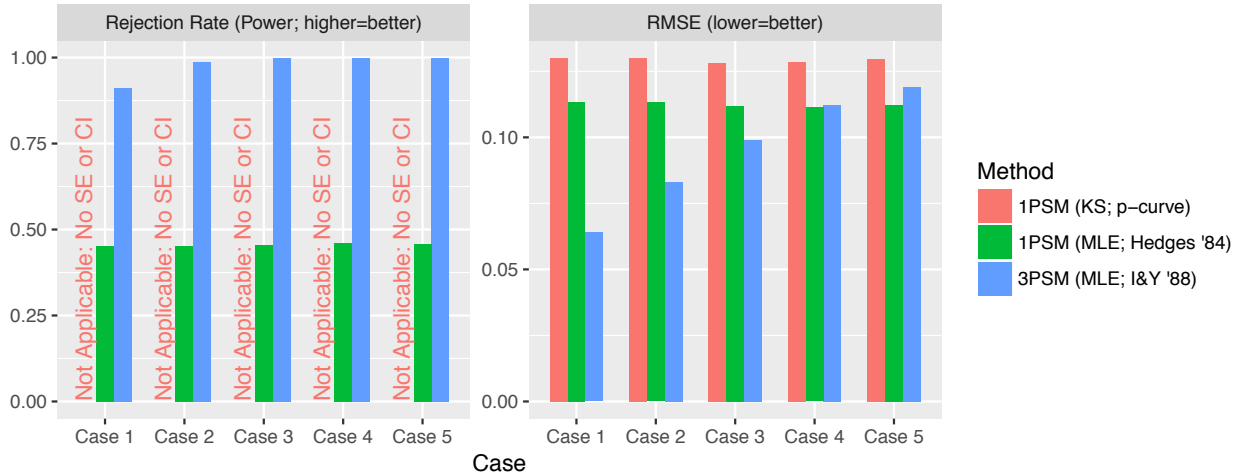

Figure 2 Small Homogeneous Effect Size. The 3PSM has much higher power and better estimation accuracy as compared to the two 1PSM implementations.

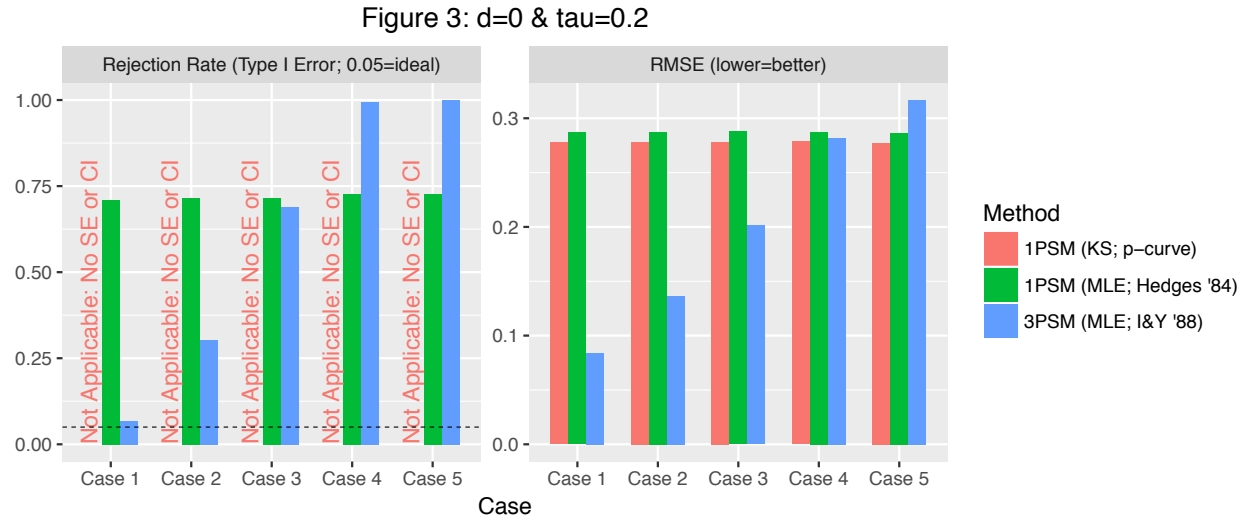

Figure 3 Zero Heterogeneous Effect Size. The 1PSM has uncalibrated Type I error for all cases while the 3PSM remains calibrated in Case 1. In terms of estimation accuracy, the 3PSM is vastly superior to the two 1PSM implementations in some cases and approximately equivalent to them in the remaining ones.

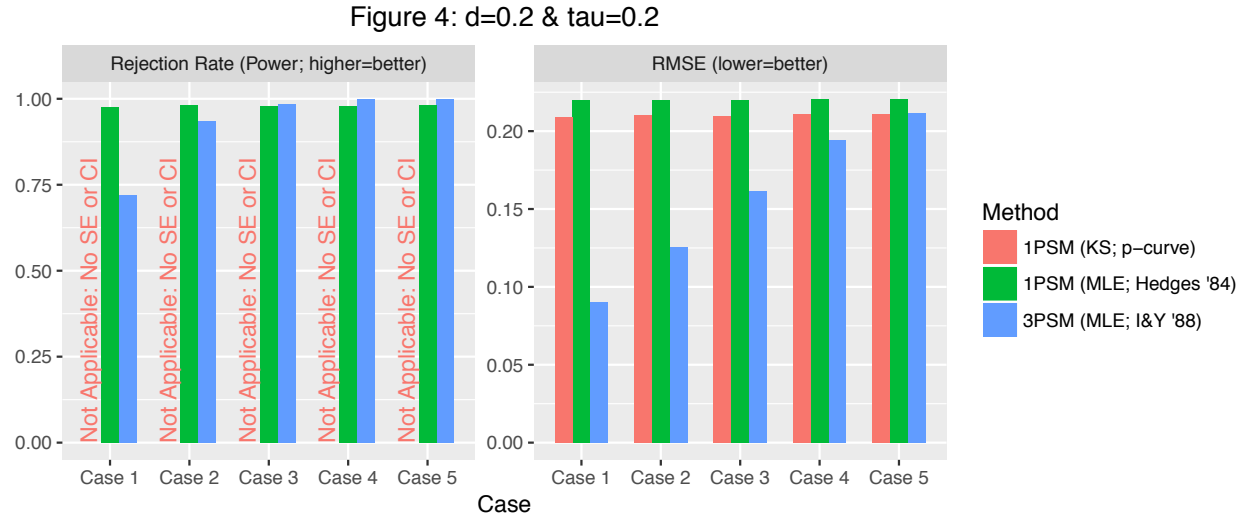

Figure 4 Small Heterogeneous Effect Size. The 3PSM generally has similar power and better estimation accuracy as compared to the two 1PSM implementations; indeed, only in Case 1 does the 1PSM have better power but this comes at the expense of highly inaccurate estimates.

References

[H1984] 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.

[MBH2016] McShane, B.B., Böckenholt, U., and 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.

[SNS2014] Simonsohn,U., Nelson, L.D. and Simmons, J.P. (2014) "*p*-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Result", *Psychological Science*, 2014, Vol.9(6), 666-681.