



On the Poor Statistical Properties of the *P*-Curve Meta-Analytic Procedure

Richard D. Morey & Clinton P. Davis-Stober

To cite this article: Richard D. Morey & Clinton P. Davis-Stober (20 Oct 2025): On the Poor Statistical Properties of the *P*-Curve Meta-Analytic Procedure, Journal of the American Statistical Association, DOI: [10.1080/01621459.2025.2544397](https://doi.org/10.1080/01621459.2025.2544397)

To link to this article: <https://doi.org/10.1080/01621459.2025.2544397>



© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 20 Oct 2025.



[Submit your article to this journal](#)



Article views: 3915



[View related articles](#)



[View Crossmark data](#)



Citing articles: 1 [View citing articles](#)

On the Poor Statistical Properties of the *P*-Curve Meta-Analytic Procedure

Richard D. Morey^a and Clinton P. Davis-Stober^{b,c}

^aSchool of Psychology, Cardiff University, Cardiff, UK; ^bDepartment of Psychological Sciences, University of Missouri, Columbia, MO; ^cMU Institute for Data Science and Informatics, University of Missouri, Columbia, MO

ABSTRACT

The *P*-curve is a widely used suite of meta-analytic tests advertised for detecting problems in sets of studies. They are based on nonparametric combinations of *p* values (e.g., Marden) across significant ($p < .05$) studies and are variously claimed to detect “evidential value,” “lack of evidential value,” and “left skew” in *p* values. We show that these tests do not have the properties ascribed to them. Moreover, they fail basic desiderata for tests, including admissibility and monotonicity. In light of these serious problems, we recommend against the use of the *P*-curve tests. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY

Received February 2025
Accepted July 2025

KEYWORDS

Admissibility; Inference;
Meta-analysis

1. Introduction



Forensic meta-analysis methods are designed to assess the quality of the evidence in sets of results. Unlike typical meta-analysis, the aim is not substantive; rather, forensic meta-analysis questions the evidence itself. Simonsohn, Nelson, and Simmons (2014a) and Simonsohn, Simmons, and Nelson (2015) describe *P*-curve analysis, a suite of statistical tests to evaluate the “evidential value,” or lack thereof, of collections of studies, potentially in the presence of poor researcher behaviors. These poor behaviors—collectively referred to as *p*-hacking (Simmons, Nelson, and Simonsohn 2011)—include selective reporting, post hoc data exclusion, and related data manipulation carried out with an aim of producing statistical significance ($p < 0.05$) that otherwise would not have been found.

The *P*-curve method has been widely applied to assess the evidence for individual phenomena (e.g., Simmons and Simonsohn 2017), collections of related empirical phenomena (e.g., Cadario and Chandon 2020; Hosseini-Kamkar, Lowe, and Morton 2021), and even whole sub-fields, such as experimental philosophy (Stuart, Colaço, and Machery 2019). Breaking out of the academic arena, the *P*-curve played a major role in a New York Times Magazine story: Dominus (2017) outlines how Simmons and Simonsohn (2017) claim the evidence for a “power posing” effect—whereby adopting an expansive physical posture increases risk taking and testosterone levels, and decreases cortisol levels (Carney, Cuddy, and Yap 2010)—is lacking, based on a *P*-curve analysis of 34 studies. In response, Cuddy, Schultz, and Fosse (2018) carried out another *P*-curve analysis on a larger set of 55 studies including the original 34 studies and found “clear” evidential value. The *P*-curve method is among the most

popular of forensic statistical procedures. Given its widespread use, it is crucial that the *P*-curve procedure be sound.

P-curve analysis is motivated by the distribution of statistically significant *p* values under three possible conditions: First, under a null hypothesis where there is collectively no effect among the many *p* values from studies being evaluated, the *p* values will often be uniformly distributed on the $[0, .05]$ interval.¹ Second, if the null hypothesis is false, *p* values will tend to gather around 0 (what the authors dub “right-skew”). Finally, if the null hypothesis is true and the researchers who carried out the original tests engaged in *p*-hacking or similar behavior to lower the reported *p* values across the threshold of significance (e.g., $\alpha = 0.05$) then the distribution of significant *p* values may tend to gather just below the significance threshold (“left-skew”). The *P*-curve statistical methodology is carried out in an attempt to adjudicate between these three conditions.

Others have criticized the *P*-curve method of Simonsohn, Nelson, and Simmons (2014a) and Simonsohn, Simmons, and Nelson (2015) from various angles. Ulrich and Miller (2015) argue that the original *P*-curve tests (Simonsohn, Nelson, and Simmons 2014a) fail to detect when researchers selectively report the outcomes of multiple tests, that is, a researcher running multiple tests but only reporting the result with the smallest *p* value; (see Simonsohn, Simmons, and Nelson 2015). Erdfelder and Heck (2019) present conditions under which right-skewed distributions of *p* values are not diagnostic of evidential value, and conservative reporting practices under which left-skewed distributions of *p* values are not diagnostic of *p*-hacking behavior; see also Bishop and Thompson (2016). Finally, Montoya, Kershaw, and Jurgens (2024) show that when there are multiple choices of which *p* value from a study to enter

CONTACT Richard D. Morey  moreyr@cardiff.ac.uk  School of Psychology, Cardiff University, 70 Park Place, Cardiff, CF10 3AT, UK

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

¹It should be noted that the condition for a valid *p* value is not uniformity, but that the *p* values are uniform or stochastically dominate a uniform. The *P*-curve authors assume that tests are not conservative.

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

into a P -curve analysis, this choice has an impact on the analysis, despite it often being arbitrary.

These critiques are important. Our approach in critiquing P -curve analysis, however, is different: we examine the soundness of the statistical procedure itself. We follow two broad lines of argumentation. In [Section 3](#), we demonstrate that the statistical tests comprising P -curve analysis are logically disconnected from the arguments presented in Simonsohn, Nelson, and Simmons (2014a) and Simonsohn, Simmons, and Nelson (2015), which leads to poor test behavior and unsupported claims. In the second line of argumentation we show that *even if* the general interpretation of the tests were sensible, the tests have poor statistical properties. These poor properties include extreme sensitivity ([Section 4.1.1](#)), test inadmissibility ([Section 4.1.2](#)), and, perhaps most disturbing, nonmonotonicity in the evidence ([Section 4.1.3](#)).

2. What is the P -curve?

There are two sets of methods Simonsohn et al. have called “ P -curve”: one set comprises meta-analytic significance tests, which we discuss in this manuscript. Simonsohn, Nelson, and Simmons (2014b) also developed methods for estimating effect sizes which we will only mention in passing (though they are related to their method for estimating “average power”, which we discuss later). For a critique of the second set of P -curve methods, see McShane, Böckenholt, and Hansen (2016). Counterintuitively, in the context of the meta-analytic tests “ P -curve” does not refer to the distribution of p values (see Ulrich and Miller 2018); “ P -curve” refers to tests based on *sums* of transformed p values.

The basic model underlying the P -curve procedure is motivated by an assumption that selective reporting—in particular, the difficulty of publishing statistically nonsignificant findings—distorts the record of nonsignificant results in an unknown way, but that significant results provide for “unbiased inferences” (Simonsohn, Nelson, and Simmons 2014a, p. 535). Thus, the first step in P -curve analysis is to collect test statistics for meta-analysis and then discard all test statistics that yield $p > \alpha_{pc}$, where the publication criterion (pc) is $\alpha_{pc} = 0.05$ by default. The statistically significant test statistics are treated as coming from a truncated distribution, where the truncation point is determined by the publication criterion.

Simonsohn, Nelson, and Simmons (2014a) and Simonsohn, Simmons, and Nelson (2015) refer to three tests: a “test of evidential value/right skew”², a “test of lack of evidential value”, and a “test of left skew”. We refer to the 2014 version of these tests as EV , LEV , and LS ; the 2015 versions, we refer to as EV^* , LEV^* , and LS^* . The distinction between them is meta-analytic, which we describe shortly.

The P -curve method is defined for a variety of common test statistics (e.g., z , χ^2 , t , F , r). Two-tailed z or t statistics are squared to yield χ^2 or F statistics, respectively. Two-tailed Pearson correlations are also converted to F statistics. To demon-

strate the method, we could choose either χ^2 or F statistics. We use χ^2 test statistics for simplicity and because they will be familiar due to the link with the standard Normal. F statistics have similar issues to the ones we describe ([Section 4.1.2](#)).

Consider a random variable X with a truncated noncentral χ^2_1 distribution:

$$X \sim \text{noncentral } \chi^2_1(\lambda, t_\alpha),$$

where λ is the noncentrality parameter and t_α is a left truncation value (e.g., the critical value for the test of $\lambda = 0$, $F_{\chi^2_1}^{-1}(1 - \alpha_{pc}) \approx 3.84$ for $\alpha_{pc} = 0.05$ where F is a cumulative distribution function). EV and EV^* test the null hypothesis that $\lambda = 0$ against the alternative that $\lambda > 0$. Simonsohn, Nelson, and Simmons (2014a) interpret a rejection as indicating that “selective reporting [is] ruled out as sole explanation for findings”/“evidential value” (p. 535).

Tests LEV and LEV^* test the null hypothesis that $\lambda \geq \lambda'$, where λ' is chosen as the noncentrality parameter that would lead to a 1/3 probability of statistical significance. For χ^2_1 test statistics, $\lambda' = 2.34$. The 1/3 probability of significance is an arbitrary value selected by Simonsohn, Nelson, and Simmons (2014a), described as “merely a suggestion” (p. 538). The value of 1/3 is hard-coded in their online app and cannot be changed. In reviewing papers that cited Simonsohn et al. (2015), we did not encounter any authors that opted for a different value.

Test LS is identical to test LEV except that it tests the less arbitrary null hypothesis that $\lambda \geq 0$; that is, it will reject when the observed test statistic is small under every possible noncentrality parameter. It is thus a test of the fit of the whole model. The analogous test LS^* was never developed by the authors and test LS was dropped without explanation from their online app. The properties of the tests will be similar to LEV/LEV^* , so we will not focus on these tests directly (but see the supplement for additional commentary).

Let p_e and p_n refer to the p values for tests EV and LEV respectively (“ e ” for “evidential value”, “ n ” for “no [lack of] evidential value”) for a single study. Here $p_e = p/\alpha_{pc}$ is simply the p value from the original test renormalized by α_{pc} to account for the truncation, and is small when X is large. p_n is the probability of obtaining a larger p value than the one observed, but still less than α_{pc} , given noncentrality parameter λ' : $p_n = (F_{\chi^2_1(\lambda')}(X) - F_{\chi^2_1(\lambda')}(t_\alpha))/(1 - F_{\chi^2_1(\lambda')}(t_\alpha))$. p_n is small when X is near the significance criterion, approaching 0 at α_{pc} .

Tests EV and EV^* dramatically increase the evidence required to argue for an effect: a single study’s results must be surprising even among significant studies ($p < \alpha_{pc}^2$) in order to be said to have “evidential value”. This is the natural effect of taking statistical significance for granted due an extreme file-drawer effect. On the other hand, p values near the significance criterion are potentially flagged as problematic by the remaining tests.

Consider a sequence of statistically significant study results X_i , $i = 1, \dots, K_\alpha$ modeled as independent truncated χ^2_1 variates with noncentrality parameters λ_i , and let K_α be the number of results that would be significant at level α . We drop α by default to mean all results significant at the conventional $\alpha = 0.05$ level. Define p values $p_{e,i}$, $p_{n,i}$ analogously as X_i .

Simonsohn et al. use two methods to combine the p values from the individual tests into a single test statistic. In 2014 they

²When Simonsohn et al. refer to “right skew”, they mean the skew of the “true” distribution of p values implied by underlying distributions’ noncentrality parameters. This will be uniform when the null hypothesis is true and right-skewed when the null hypothesis is false under many, but not all, common tests. See Gelman and O’Rourke (2014) for further critical discussion.

use Fisher's method of summing the logarithms of p values (as do van Assen, van Aert, and Wicherts (2015), who suggest test EV at around the same time); in 2015 they use Stouffer's method of summing probit-transformed p values (see Marden (1985) p. 1536 for a description of various meta-analytic methods for combining p values). Under the null hypothesis $\lambda_i = 0, \forall i$ the resulting sums have central χ^2_{2K} and Normal distributions, respectively; p values are computed relative to these null distributions:

$$p_{e,\cdot} = 1 - F_{\chi^2_{2K\alpha}} \left(-2 \sum_{i=1}^{K_\alpha} \log(p_{e,i}) \right)$$

(Simonsohn, Nelson, and Simmons 2014a)

$$p_{e,\cdot}^* = \Phi \left(\frac{1}{\sqrt{K_\alpha}} \sum_{i=1}^{K_\alpha} \Phi^{-1}(p_{e,i}) \right)$$

(Simonsohn, Simmons, and Nelson 2015)

where Φ and Φ^{-1} denote, respectively, the cumulative distribution function of the univariate normal distribution and its inverse. We can define $p_{n,\cdot}$ and $p_{n,\cdot}^*$ analogously. For demonstration purposes, see our interactive app that performs the P -curve tests at <https://richarddmorey.github.io/pcurveAppTest/>.

For the meta-analytic tests EV and EV^* the null hypothesis tested is that $\lambda_i = 0$ in all studies; hence, the alternative is that $\lambda_i > 0$ for at least one i . For LEV and LEV^* , the null hypothesis is that all $\lambda_i \geq \lambda'$; hence, a rejection would typically be interpreted as favoring the alternative $\lambda_i < \lambda'$ for at least one i .

Simonsohn, Simmons, and Nelson's (2015) change from summing logarithms (2014) to summing probits (2015) was accompanied by a change in their online app (Simonsohn, Nelson, and Simmons 2017), which no longer offers the 2014 tests. Citing Abelson (1995), their justification was that the probit combination method is "less sensitive to a few extreme results" (p. 1149), a property they preferred. This choice has dramatic implications for the statistical properties of the method, to which we turn later.

A second change that Simonsohn, Simmons, and Nelson (2015) made was to define a procedure including the "full" P -curve, as described above and a "half" P -curve, which is defined analogously but the truncation point is $p < \alpha_{pc}/2$ instead of α_{pc} . They argue this accounts for "aggressive p hacking": that is, that some people's target significance criterion when engaging in bad behavior may be lower than 0.05. They (arbitrarily) choose half the usual criterion as a secondary criterion and build a new test procedure: "We introduce the following novel test of evidential value: A set of studies is said to contain evidential value if either the half P -curve has a $p < 0.05$ right-skew test, or both the full and half P -curves have $p < 0.1$ right-skew tests." (Simonsohn, Simmons, and Nelson 2015, p. 1151, emphasis in original) For most of this article we do not use this compound test procedure, instead concentrating on each test on its own. We return to the compound criterion later to show that it leads to nonmonotonicity in the evidence.

2.0.1. Use of the Methods in the Literature

As of July 2024, Simonsohn, Nelson, and Simmons (2014a; log method) and Simonsohn, Simmons, and Nelson (2015; probit

Table 1. Cross tabulation of the results of P -curve analyses in papers that cite Simonsohn, Simmons, and Nelson (2015).

	Test	N	%	K	
				Med.	Range
$p \geq 0.05$	$p \geq 0.05$	25	9.3%	9.0	(1–22)
	$p < 0.05$	12	4.4%	17.5	(6–41)
	Did not report	5	1.9%	5.0	(3–35)
$p < 0.05$	$p \geq 0.05$	189	70.0%	19.0	(1–569)
	$p < 0.05$	2	0.7%	65.0	(60–70)
	Did not report	37	13.7%	13.0	(2–282)

method) collectively have about 1200 citations, most of which are applications of the method. We reviewed all P -curve tests in the 186 papers that cited Simonsohn, Simmons, and Nelson (2015); see Table 1 for a summary. More details can be found in the supplement. The vast majority of applications of EV^* were significant (84.4% of P -curves), and the modal conclusion from applying P -curve methodology was "evidential value" for whatever set was reported. Only rarely (4.4%) did a P -curve analysis yield a statistically nonsignificant EV^* test and a significant LEV^* test (supposedly indicating a "lack of evidential value").

3. Conceptual Issues with the P -curve

Before our formal statistical critiques of the P -curve, we offer a number of critiques centering on its conceptualization and use.

Absurd repeatable process. The repeatable process assumed by the P -curve authors (and their power analyses) uses a fixed number of significant results for all repetitions. Given that p values are random variables, it is quite strange to assume that a given number of draws from a distribution of p values will yield *exactly* the same number of p values less than 0.05 (both p -hacked and non- p -hacked) each time. This naturally leads to absurdities. If we take this assumption at face value, then researchers would also be p -hacking to nonsignificance (i.e., above 0.05) to maintain a fixed number of significant p values across repetitions. The important critique here is that the P -curve authors do not consider plausible models of researcher behavior and hence ignore an important source of variability.

Not tests of skew. Simonsohn, Nelson, and Simmons (2014a) and Simonsohn, Simmons, and Nelson (2015) describe their tests in terms of the skew of the observed p values (i.e., " P -curves that are not right-skewed suggest that the set of findings lacks evidential value," Simonsohn, Nelson, and Simmons 2014a p. 535). However, the P -curve tests are all based on sums of transformed p values. There are many sets consistent with a given sum: for instance, any sum of K values is consistent with a set of K identical values. This set has no skew. Manipulation of this set of identical values can yield sets of left or right skew with the same sum. The test is only responsive to the size of the sum, not the skew of the values within it. It is true that highly right-skewed samples of p values often lead to rejections of EV^* , but left-skewed samples can do the same.

Given that the theoretical distribution of p values when at least one $\lambda_i > 0$ is right-skewed (assuming no poor researcher behavior), one might be tempted to consider a test of all $\lambda_i = 0$ with the alternative at least one $\lambda_i > 0$ a test of skew. In this case, there is nothing special about skew; we could just as easily regard the P -curve tests as tests of the expected p value (because

the expected value of the p values decrease as some λ_i increases), or the variance (which decreases as some λ_i increases).

Moreover, under the assumptions of Simonsohn, Nelson, and Simmons (2014a) and Simonsohn, Simmons, and Nelson (2015), there are no effect sizes that would lead to left skew (i.e., a preponderance of p values near the publication threshold). Thus, tests of λ cannot be tests of left skew, despite that they advertise a test of left skew (LS).

Taken together, the authors' use of the concept of skew is inconsistent with both sample skew and population skew. The P -curve tests are simple tests of noncentrality parameters; introducing the concept of skew obfuscates how the tests work.

Noncentrality parameters, not effect sizes. Tests of noncentrality parameters are not equivalent to tests of effect sizes. Given that the input to the P -curve procedures is a set of p values, and one persistent critique of p values is that they confound effect size and sample size (e.g., Berkson 1938, onward), one might wonder how it is that the P -curve tests can be tests of effect size. In truth, they are not.

When testing the null hypothesis that the effect size is 0, we can elide the difference because when the noncentrality parameters are 0, so are the effect sizes. Tests EV and EV^* are tests of whether the average transformed p value is what would be expected if all considered effect sizes are exactly zero, with an alternative that at least one noncentrality parameter $\lambda_i > 0$.

Tests LEV and LEV^* , on the other hand, are tests of whether the average transformed p value is what would be expected if the χ^2_1 noncentrality parameters, from all considered studies, are all larger than 2.34 (*mutatis mutandis* for other test statistics). In a typical two-tailed z test (performed as a χ^2 test from which the X_i values may derive) the noncentrality parameters are $\lambda_i = \delta_i^2 N_i$, where δ_i and N_i are standardized effect size and effective sample size, respectively. The hypothesis being tested by LEV and LEV^* is dependent on the design of the studies at hand: they are tests whose alternative hypothesis is that “for at least one study i , $|\delta_i| < \sqrt{2.34/N_i}$.” This is in contrast to Simonsohn, Nelson, and Simmons's (2014a) claim that LEV , for instance, is a test “not that the effect is zero, but that it is very small instead” (p. 537). Conflating effect size with noncentrality parameters obfuscates what the hypotheses actually are; moreover, there is not a single effect size as implied by “the effect size”: each study will have its own noncentrality parameter.

Not about the “power” in a set. Simonsohn, Nelson, and Simmons (2014a) claim that tests LEV and LEV^* are “test[s] for power of 33%” (p. 537). In a companion paper, Simonsohn, Nelson, and Simmons (2014b) suggest that that the “ P -curve can be used to estimate the average underlying statistical power of a set of studies” (p. 676; though it should be noted that the P -curve analysis in that paper is slightly different, though related, to the P -curve analysis we discuss here). There are two reasons why this is not true. The first is conceptual: their “power” is a function of the true effect size, which is at odds with the classical concept of power (a function of counterfactual, not true, effect sizes). The second is statistical: their estimate is inconsistent, even if taken at face value. We address the latter critique in Section 4.3.

Logical disconnect between statistical hypotheses and “evidential value.” It is common for authors to interpret the results of the P -curve procedures in terms of the set of studies, of a broader research area, or in terms of a theoretical construct (like

a psychological “effect”). Researchers may even split a meta-analysis into subsets and use P -curve to look for where the evidential value lies (e.g., Stuart, Colaço, and Machery 2019).

When rejecting the null hypotheses of EV , EV^* the natural conclusion is that *at least one noncentrality parameter is nonzero*. Likewise, when rejecting the null hypotheses of LEV , LEV^* the natural conclusion is that *at least one noncentrality parameter is less than 2.34* (see Section 4.2). The P -curve, however, is intended to draw conclusions regarding properties of sets of studies; these null and alternative hypotheses, which still involve individual studies, are clearly logically disconnected from this purpose. For any of the six tests, the only proper conclusion to draw is that it is not the case that *all* the studies have the property in question.

If a single study in a set of 50 truly had a nonzero noncentrality parameter, would it be valid to claim that, for example, the corresponding literature from which the set is drawn is high-quality? Clearly not; however, it is common for P -curve practitioners to reject EV or EV^* and come to a similar conclusion. For instance, Wilde (2022) points to a statistically significant EV^* test and writes “We should think of this result as a quality control stamp supporting the validity of the widely-noted correlation [between jaundice and autism]” (p. 23). Ruz et al. (2020) claims that a significant EV^* test means that “this set of studies contains evidential value for reward-driven distraction.” (p. 886).

The problem is analogous for LEV and LEV^* : should a single study with a small sample size and smaller-than-expected noncentrality parameter “contaminate” an otherwise robust set of studies? For all tests, the use of a sum in computing the meta-analytic test statistic ensures this contamination will happen, but the typical conclusions based on that sum are far removed from what is warranted.

It is also worth pointing out that “evidential value” is not a well-defined statistical or scientific concept and the P -curve tests can come to contradictory conclusions regarding it. It is possible for test EV/EV^* to reject and for LEV/LEV^* to reject at the same time (see West et al. (2021) for a real example in the literature). A set surely cannot both contain evidential value and lack it.

4. Statistical Properties of the P -curve

4.1. Properties of Tests EV and EV^*

4.1.1. Undue Sensitivity at the Critical Boundary

Consider a hypothetical set of three studies, where each study reports the results of a single z test as follows: $z_1 = 4$; $z_2 = 3.4$; and $z_3 = 1.964$. If we test the “evidential value” for these studies using EV^* , we obtain $p_{e^*}^* = 0.048$. Therefore, we would conclude that this set “has evidential value”. However, suppose that the authors of the original study 3 had decided to round z_3 to two digits instead of three, to be in line with a journal's style requirements. For a well-behaved meta-analytic procedure, this should produce a trivial change (if any) in the overall test. If $z_3 = 1.96$, however, test EV^* yields $p_{e^*}^* = 0.198$. A small change in the statistic—bringing it closer to the boundary through rounding—has caused a major change in the p value and a difference in the conclusion at the standard level $\alpha = 0.05$.

This sensitivity is the result of how the “robustness” was achieved in the 2015 switch to the probit in the P -curve tests.

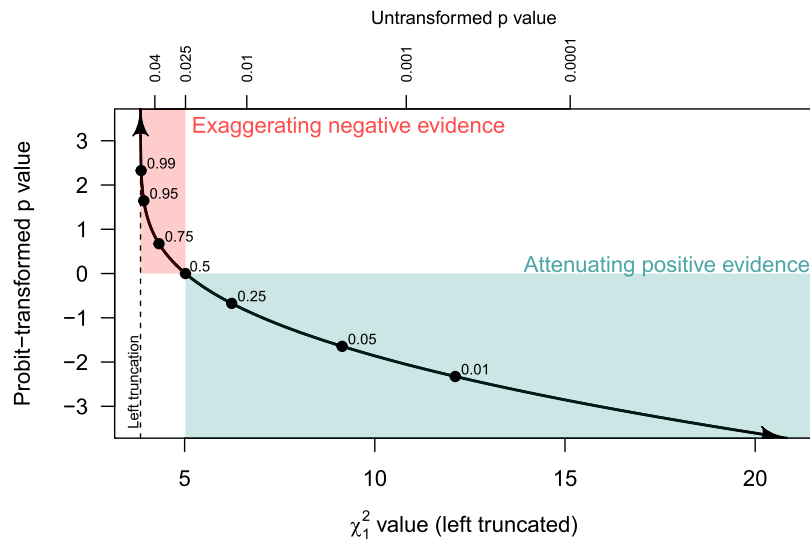


Figure 1. Transformation from the truncated χ_1^2 values to probit-transformed values that will be summed to produce $p_{e^*}^*$ in test EV^* . The points show the p values from the χ_1^2 conditional on being above the left truncation point.

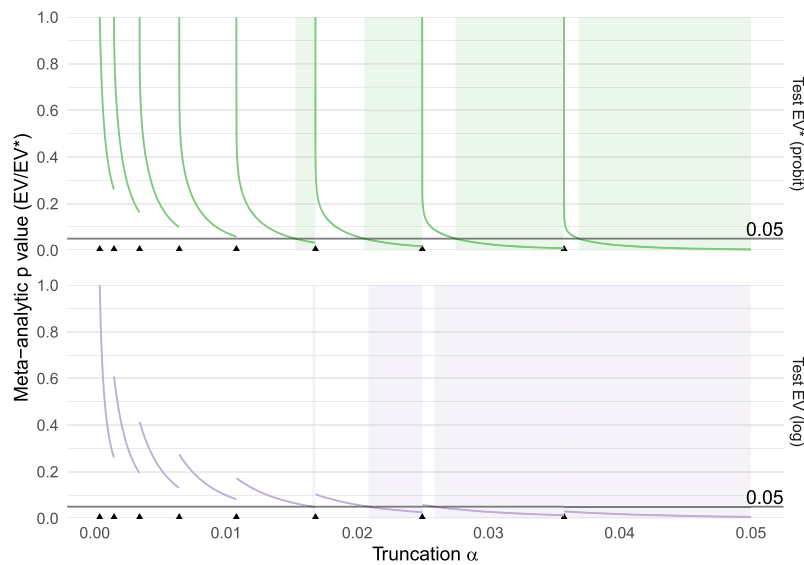


Figure 2. Meta-analytic p values as the truncation bound is changed. Top: Test EV^* ; Bottom: Test EV . Shaded regions show the truncation points where the $p_{e^*}^*$ (top) or p_{e^*} (bottom) is statistically significant. Small, black triangles at the bottom of each plot show the location the 8 p values entered into the meta-analysis.

Figure 1 shows the transformation from the truncated χ_1^2 test statistics to the values to be summed for test EV^* . Small χ_1^2 values near the truncation point are transformed to arbitrarily large values. This has the effect of exaggerating the evidence *against* an effect when they are summed with other values. On the other hand, large values of the test statistic are increasingly attenuated by the transform.

If we could summarize the main error in choosing the probit transform for this problem, it is that the probit is a two-tailed transformation (strongest evidence at $-\infty$ and $+\infty$), but test EV^* applies it to one-tailed test statistics (strongest evidence at $+\infty$).

We can view the sensitivity as an interaction between the p values in a set and any criterion. Recall that Simonsohn, Simmons, and Nelson (2015) proposed the “half P -curve” rule, which combines a test where the truncation occurs at $p = 0.05$ with one where the truncation occurs at $p = 0.025$. The probit

transform causes test EV^* to be poorly behaved, because a value near *either* criterion can completely negate all other evidence. The second criterion also adds arbitrariness to the test. The reason for $\alpha_{pc} = 0.05$ is clear: it is the standard threshold for statistical significance. The same cannot be said for the half P -curve’s $\alpha_{pc} = 0.025$, which was chosen arbitrarily. The test result can change radically depending on which criterion we happen to choose for the “half” criterion. In fact, we can almost achieve any test outcome we like.

To show this, we produced a set of eight studies whose individual p values were 9-iles of the distribution of statistically significant p values when $\lambda = 3.84$ (i.e., the noncentrality parameter yielded a 0.5 probability of significance in any single study). Figure 2 shows test EV^* ’s $p_{e^*}^*$ (top) and EV ’s p_{e^*} (bottom) as the truncation point changes.

Figure 2 (top) shows what happens for the constructed dataset as we change the second criterion (the “half” criterion).

For test EV^* a single result near the criterion produces arbitrarily large p values. As the significance criterion drops, $p_{e_i}^*$ increases to 1. When a study is then eliminated ($p_i \geq \alpha_{pc}$), $p_{e_i}^*$ drops from 1 to a much lower value. This is shown in Figure 2 (top) as lines that swing wildly between 0 and 1. This value of $p_{e_i}^*$ is almost entirely dependent on which α_{pc} we choose for the “half” test. Although Simonsohn, Simmons, and Nelson (2015) chose 0.025, any choice of α_{pc} between 0.02 and 0.04 could represent “ambitious” targets for p hacking and would be defensible for the half P -curve’s purpose. If the set contains even a single p value in that range, there is a choice of criterion that will negate all other evidence. Our hypothetical dataset contains a study with $p = 0.02495$ that is close to the half P -curve’s sharp discontinuity at 0.025, and thus is in one of the wild upward swings ($p_{e_i}^* = 0.193$).

Test EV is much less sensitive. Because the central χ_1^2 distribution has an exponential right tail, the test EV ’s p value will be approximately proportional to $e^{-X/2}$ and hence $\log p_i \propto X$. Because the transformation used for test EV is almost linear in the test statistic, there is little sensitivity just above the lower bound and no attenuation as the test statistic grows.

In Figure 2 (bottom), we can see mild discontinuities in test EV ’s p value when a study crosses the “half” P -curve boundary and moves from being within the set to outside the set. When combining logarithms, a value at the criterion has $\log(1) = 0$ and hence does not contribute to the sum. However, when the study is dropped ($p_i \geq \alpha_{pc}$) the degrees of freedom of the test decrease so the significance criterion changes. Dropping the “dead weight” study near the criterion decreases p_{e_i} slightly. As the truncation α_{pc} becomes very low and few studies are left in the set, the changes in degrees of freedom are more substantial and hence the differences are larger.

4.1.2. Inadmissibility

Although the sensitivity we outlined in the previous section looks objectionable, it raises a question: *how much sensitivity is too much?* Simonsohn, Simmons, and Nelson’s (2015) changed from the log to the probit transformation simply because they preferred the answers from the probit transformation. This arbitrariness is objectionable, but if we merely object to the sensitivity, we have not done much better. We must examine the tests more closely to understand whether the sensitivity is a symptom of some deeper issue with how the test is using the data.

If we assume under the null hypothesis that p values have a uniform distribution, then we can choose any inverse CDF as a transformation; under the null, the transformation will have the corresponding distribution. Choosing a family of distributions that is closed under summation, such as χ^2 (the log method) or Normal (the probit method) ensures that we can easily determine the null distribution. Thus, in the applied meta-analytic literature, combination procedures are chosen mostly for this particular convenience, along with heuristics about how these methods weight p values (Abelson 1995).

Since Birnbaum (1954; see also Matthes and Truax 1967), however, it has been known that there is much more to the choice of transformation for meta-analytic combinations. Understanding the distribution under the null hypothesis is not enough; a transformation must also be tailored to the test statistics in

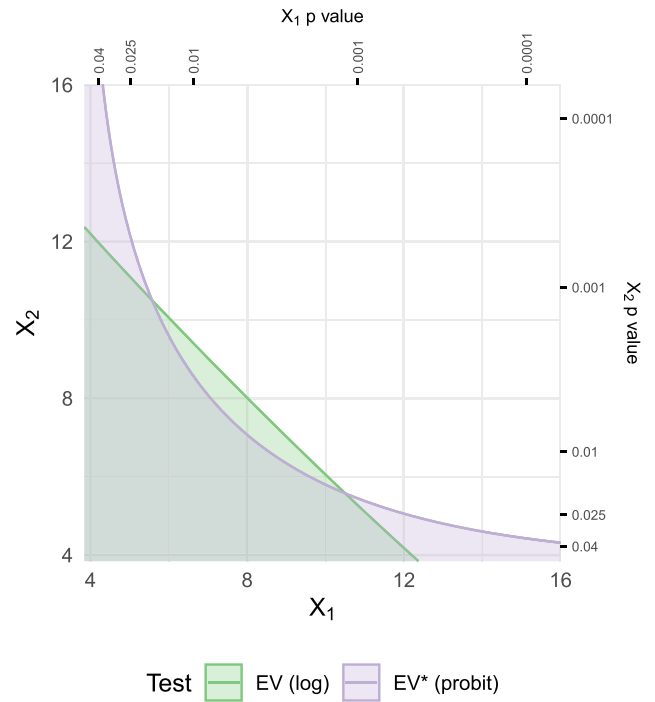


Figure 3. Acceptance regions for tests EV and EV^* when $K = 2$. Note that the acceptance region for test EV^* is concave.

a way that respects their distribution under alternative non-centrality parameters. Some transformations yield inadmissible tests: their power function is dominated by some other choice of transformation for *all* alternatives. Birnbaum showed that for exponential families, a necessary and sufficient condition for admissibility is convexity of the acceptance region in the test statistic. Marden (1982) later proved that convexity is also necessary and sufficient for noncentral χ^2 distributions, and we extend his proof to truncated noncentral χ^2 distributions as used by the P -curve procedures (see supplement).

Figure 3 shows the acceptance regions for tests EV and EV^* . Test EV^* ’s acceptance region is concave, not convex, and must be so for all K owing to the probit transform. Because $\lim_{p \rightarrow 1} \Phi^{-1}(p) = \infty$, a value of p_i close enough to α_{pc} can always negate *any amount* of evidence from the other $K - 1$ studies. Indeed, this issue is in play in Section 4.2.2; even more extreme examples could be constructed with a single suitably small p value. The acceptance region has nonzero volume along each of the k rays from the “origin” ($t_\alpha, \dots, t_\alpha$) and cannot be convex for any K .

Test EV ’s acceptance region, on the other hand, will be approximately bounded by a simplex due to the exponential tail of the χ_1^2 distribution. Adapting Marden’s proof to the P -curve procedures reveals that the probit transform yields an inadmissible test when combining χ^2 test statistics, but the log transform is admissible.

One of the properties of a convex acceptance region is that a single large value in any of the X_i values can guarantee a rejection. Test EV^* , though, was specifically engineered to be “robust to extreme results” through the switch to the probit transform (Simonsohn, Simmons, and Nelson 2015, p. 1149). But any test with a convex acceptance region—any *admissible* test—will have precisely the property that Simonsohn et al. object to: sensitivity

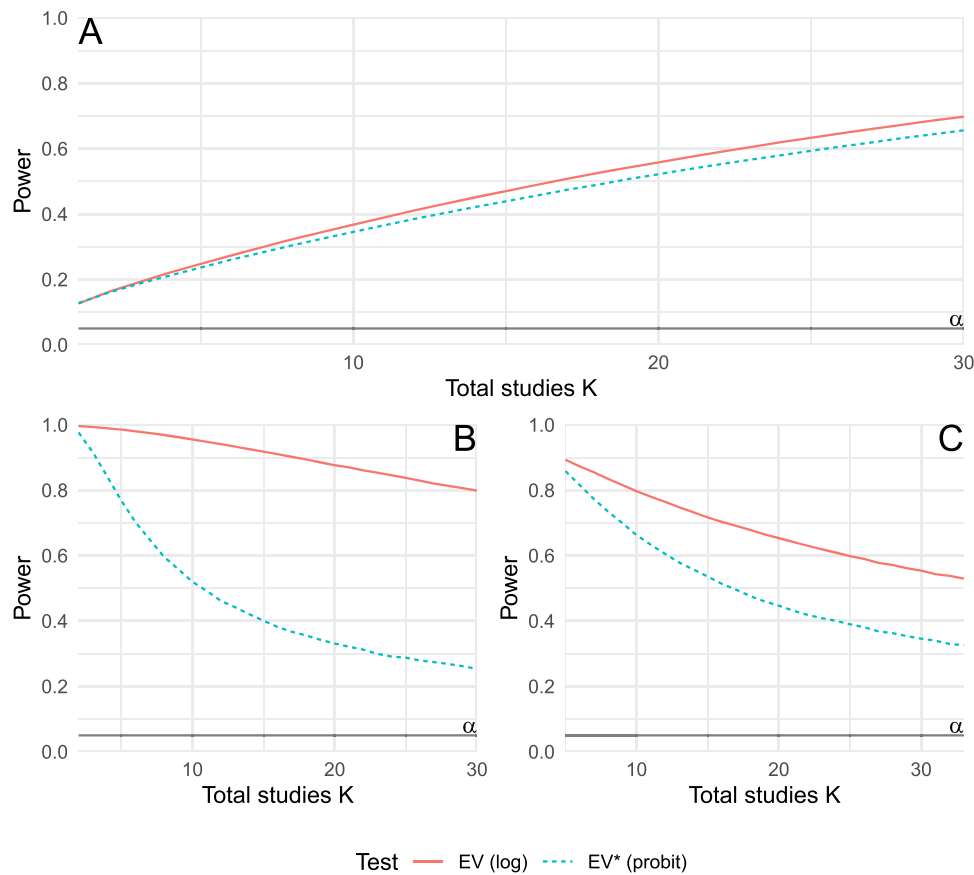


Figure 4. Power of tests EV and EV*. A: All K studies $\lambda = 1$. B: $K - 1$ studies have $\lambda = 0$, and 1 study has $\lambda = 36$ (equivalent to $\delta = 0.6$ if $N_{\text{eff}} = 100$). C: $K - 4$ studies have $\lambda = 0$, and 4 studies have $\lambda = 8$ (equivalent to $\delta = 0.28$ if $N_{\text{eff}} = 100$).

to large, individual X_i values. All symmetric admissible tests of size α must reject for values when one X_i is beyond some minimum value. In engineering out a reasonable property of the test, they have produced a test EV^* that is inadmissible and inappropriately sensitive to *small* values of X_i . Indeed, it is hard to see why a test of the hypothesis that “all $\lambda_i = 0$ ” should *not* reject if one of the test statistics is large enough.

Unfortunately, simply knowing that a test is inadmissible does not tell us which other test dominates it; it also does not tell us how the test will perform against other tests, including admissible ones. It is not the case that every admissible test will dominate every inadmissible test. We should therefore compare power curves across tests to see when, in particular, the inadmissible test might be failing to capture the information in the data.

Inspection of the acceptance regions in Figure 3 and previous work with non-truncated χ^2 variables (Kozioł and Perlman 1978) suggests that the best case scenario for test EV^* will be when all studies have the same noncentrality parameter. If the studies are heterogeneous, on the other hand, the power of test EV^* would be expected to suffer because its acceptance region over-weights test statistics near the significance criterion. Figure 4(A) and (B) show power curves that confirm this intuition. In all cases, the power for test EV^* is lower than test EV ’s; in the heterogeneous case, test EV^* ’s power is dramatically lower. Figure 4(C) shows that the power reduction is still substantial when several studies (in this case, four) have a moderate effect size and the others are null.

We want to stress, however, that our main complaint is not about low power per se: rather, the lower power from inadmissibility is a sign that the test is not using information in the data regarding the hypotheses being tested. Power is an average property of a test; we would also like for a test to yield reasonable results for particular datasets we might observe. Test EV^* has had a reasonable property—the sensitivity to a few large values—engineered out, and thus its power suffers against other tests even in cases where there are *not* individual large values. The probit transform exaggerates negative evidence and attenuates positive evidence, so it is not a surprise that its power suffers.

Extension to F statistics. Simonsohn, Nelson, and Simmons (2014a) and Simonsohn, Simmons, and Nelson (2015) square t statistics to obtain F_{1,v_2} test statistics, and transform Pearson correlations to t test statistics that are then squared. Marden (1982) showed that for noncentral F statistics, probit combinations are always inadmissible (see supplement for our extension to truncated F statistics). He also showed, however, that Fisher’s method of combining logarithms is inadmissible with noncentral F statistics when the numerator degrees of freedom are less than 2, which rules out squared t statistics. This implies that for some α_{pc} , test EV will also be inadmissible; however, admissibility conditions for EV with t statistics are not known. We show the dramatic power implications for F statistics analogous to the ones for χ^2 statistics in the supplement.

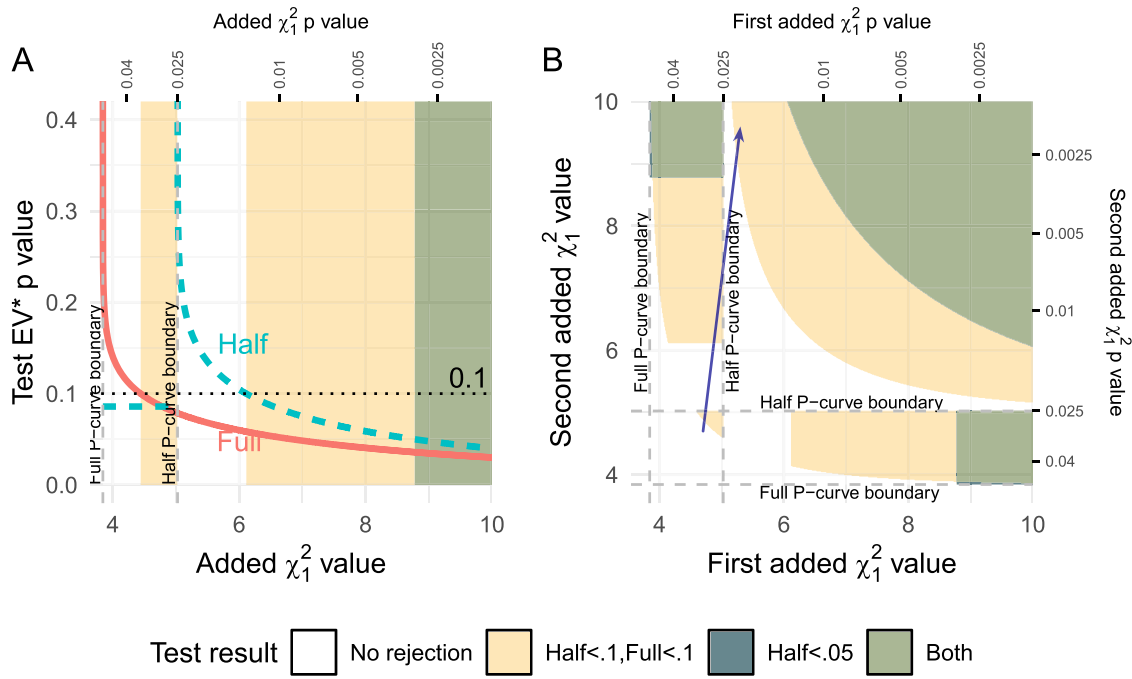


Figure 5. Results of Simonsohn et al.'s combined procedure with the probit method, increasing a single test statistic (A) and two test statistics (B) added to the demonstration dataset. The shaded regions show when would be statistically significant, and why. In B, the arrows show paths through the space such that both data values increase, yet the decision can change from rejection to acceptance several times.

The attraction of particular transformations is that they are supposed to be “nonparametric” (Marden 1982) and thus widely applicable to many underlying test statistics. But if different test statistics require different transformations prior to summing to respect the evidence, this raises the question of whether one should combine them as the P -curve does. Most of the statistical literature on meta-analytic combination methods focuses on summing statistics in the same family, not *across* families. More theoretical work in this area is necessary before one can have confidence when combining across families.

4.1.3. Nonmonotonicity in the Evidence

One property that is generally considered essential for a reasonable statistical test is *monotonicity in the evidence*. Increasing the amount of evidence in a single data point should never cause a test to move from rejection to acceptance. It is such a reasonable and widely-met condition, in fact, that it led (1954) to search for other ways to distinguish meta-analytic combination methods, such as admissibility.

Simonsohn, Simmons, and Nelson’s (2015) compound half P -curve procedure creates a test that is non-monotone in the evidence, due to the hard boundaries between the full and half P -curve. The half P -curve adds a second boundary in the middle of the test statistic space. When a test statistic lies just on the smaller side of this boundary, it contributes to the strength of the evidence in the full P -curve but not at all to the half P -curve. If the test statistic is increased slightly to be just on the larger side of the half P -curve boundary, it contributes slightly more evidence to the full P -curve, but *penalizes* the p value of the half P -curve due to its proximity to the criterion.

In the case of the probit combination method, this reduction of $p_{e.}$ for the half P -curve is extreme due to the mapping of the boundary to $+\infty$. To show this, we constructed a demonstration

dataset containing $k = 15$ χ^2_1 test statistics (values given in the supplement). We then introduced an additional data point, manipulating the test statistic from small to large.

Figure 5(A) shows how the full and half P -curve tests respond to this increase in the evidence. The shaded regions show where the test is statistically significant by Simonsohn, Simmons, and Nelson’s (2015) compound rule, and the shading color shows why. Before the new point reaches the half P -curve boundary, it can increase the evidence in the full P -curve. When the full P -curve’s $p_{e.}$ drops below 0.1, both the half and full P -curve are less than 0.1, triggering Simonsohn, Simmons, and Nelson’s (2015) compound rule.

However, when the new test statistic crosses the half P -curve boundary, all other evidence in the half P -curve is negated and the half P -curve’s $p_{e.}$ increases to 1. Thus, the compound procedure no longer indicates rejection of the null hypothesis. As the test statistic increases, it again begins to contribute evidence to the half P -curve and eventually the compound rule indicates rejection. The half P -curve criterion leads to moves from nonrejection, to rejection, back to nonrejection, then back to rejection again in spite of an steady increase in the evidence away from the significance threshold α_{pc} .

This behavior is especially strange considering the intention of the compound procedure: to help make the procedure robust to “ p hacking”. As Simonsohn, Simmons, and Nelson (2015) make clear, p hacking to smaller p values is more difficult; a nonmonotone procedure, then, penalizes a set for containing a value that is *less* consistent with p hacking when it crosses the half P -curve boundary.

Figure 5(B) shows the rejection regions of the compound test procedure when two new data points are added instead of one. As the arrow shows, it is possible to increase the two data values and change decisions five times along the way. More changes

are possible when taking into account larger numbers of data points changing, though higher-dimensional spaces are difficult to visualize.

Can the half P -curve procedure be rescued by returning to the logarithmic combination procedure, which is less sensitive at the boundary? No; the log method is less sensitive, but still produces a non-monotone test procedure due to the half P -curve boundary. We leave this demonstration to the supplement.

One could argue that the P -curve's truncation at 0.05 causes the same kind of nonmonotonicity, even without applying a compound decision rule. Imagine a set of hypothetical p values yielding a statistically significant test EV^* . Consider adding a new point to the set. A new hypothetical p value just above 0.05 is clearly some evidence, and a hypothetical p value just below 0.05 cannot represent less evidence (i.e., p hacking cannot reduce the evidence from the un- p -hacked data). A hypothetical new p value just above 0.05 would be completely ignored, leading to the same significant EV^* test as without that new value. A new hypothetical p value just below 0.05 would negate the evidence in the rest of the set, causing P -curve users to doubt the whole set. The hard threshold of the P -curve leads to a fundamental nonmonotonicity, even without their compound decision rule.

4.2. Properties of Tests LEV and LEV^*

Tests LEV and LEV^* are tests of the hypothesis that *at least one noncentrality parameter* in a set of studies is smaller than $\lambda' \approx 2.34$, after conditioning on statistical significance. The first critique—and the most substantial one, in our opinion—is that it is not clear why one would test this hypothesis. Simonsohn, Nelson, and Simmons (2014a) have interpreted this test as a test of “lack of evidential value,” implying that if this test rejects, then this means that the distribution of p values “is flatter than one would expect if studies were powered at 33%”. But there is no logical connection between the statistical hypothesis test and their interpretations.

To see this more clearly, consider that the test statistics for LEV and LEV^* are sums of transformed p values. As one p_i value in the set approaches α_{pc} , $p_{n,i}$ and $p_{n,i}^*$ will approach 0, and hence $p_{n,\cdot}$ and $p_{n,\cdot}^*$ must approach 0. This is not necessarily problematic for a test that at least one noncentrality parameter is small, but it is inconsistent with the interpretation that the whole set “lacks evidential value” or that it is about the “flatness” of the distribution of p values. The flatness of the distribution of p values will hardly change as a single p value approaches 0.05, yet the test is guaranteed to eventually reject regardless of how large the other test statistics are.

A further reason why this hypothesis is uninteresting is that noncentrality parameters confound effect size and sample size.

4.2.1. Problematic Interpretation Due to the Bounded Parameter Space

The boundedness of the parameter space at $\lambda = 0$ further complicates the interpretation of tests LEV/LEV^* . When a significance test rejects, the conclusion is that the null hypothesis misfits in some way. With an unbounded parameter space, the move from “the null misfits” to “the effect size is large” does not seem problematic because no matter how large the test

statistic is, there are parameter values consistent with it (though in practice we try to ensure rejection isn't due to misfit of another aspect of the model).

With a parameter space bounded below, however, as a test statistic shrinks it moves from supporting small effect sizes to supporting model misfit. A classic example of this logic is Fisher's famous critique of Mendel (Fisher 1936; Edwards 1986): small test statistics lead not to a conclusion that the noncentrality parameter is small, but rather that something has gone wrong. In the case of LEV/LEV^* , there is little difference between the critical p value in a single study that would lead to a rejection of $\lambda \geq 2.34$ at $\alpha = 0.05$, and the critical p value that would lead to a rejection of the model outright at $\alpha = 0.05$ (i.e., reject $\lambda \geq 0$): 0.045 and 0.0475, respectively. A rejection of the hypothesis that “all $\lambda_i \geq 2.34$ ” may not be grounds for claiming that the noncentrality parameter is small; it may only be grounds for claiming that the null model misfits. The test statistic is so consistently close to the significance boundary that we reject the hypothesis that every test statistic was an independent realization of a random variable of the sort specified by the meta-analyst.

The P -curve authors might be satisfied in concluding “model misfit” instead of “there is at least one small noncentrality parameter”: the misfit might be caused by p hacking. This conclusion, though, does weaken what one can say on the basis of a rejection of the null in tests LEV/LEV^* , particularly because the null model is “all $\lambda_i \geq 2.34$ ”. When this test rejects we might ask, “so what?” A rejection of *all* noncentrality parameters, is potentially more interesting; in the supplement we discuss misfit tests LS/LS^* (and a less-sensitive replacement).

4.2.2. Sensitivity at the Criterion

We previously showed that test EV^* was highly sensitive to one value at the criterion, but test EV was not. Test LEV , however, is very sensitive to a single value at the criterion, while test LEV^* inherits the sensitivity of EV^* .

The sensitivity of LEV is due to the effective reflection of the p values before entering the values into the log transformation, instead of simply using the same test statistic as test EV . A p value of 0.049, for instance, will yield a value to be summed of $-2 \log(0.0094) = 9.3$; considering that the expectation of each component of the sum is 2 (because each study increases the degrees of freedom by 2), such a value has a lot of weight.

For test LEV this becomes extreme: consider six studies that each with $Z = 8$ combined with a single study that yields $\chi^2_{195} = 228.58$ ($p = 0.0499997$). Although six studies have very large Z statistics, they are canceled out by the single value near the significance criterion. Adding to the problem, arbitrarily large test statistics are mapped to 0 in the sum because for large X_i values $p_{n,i} = 1$ and $\log(1) = 0$; hence, they do not contribute to the sum at all! The example thus works just as well with six studies with $Z = \infty$. To claim that a rejection of the null in test LEV indicates a “lack of evidential value” would be absurd.

Owing to the thinner tails of the probit transformation test, LEV^* is less sensitive than LEV , but it is still very sensitive. The reflection of the p value doesn't change the sensitivity from EV^* because the probit transformation is symmetric. The set $Z = \{1.964, 2.8, 2.8\}$ yields $p_{n,\cdot} = 0.178$. If an author were to round 1.964 to 1.96, this results in $p_{n,\cdot} = 0.045$. As with

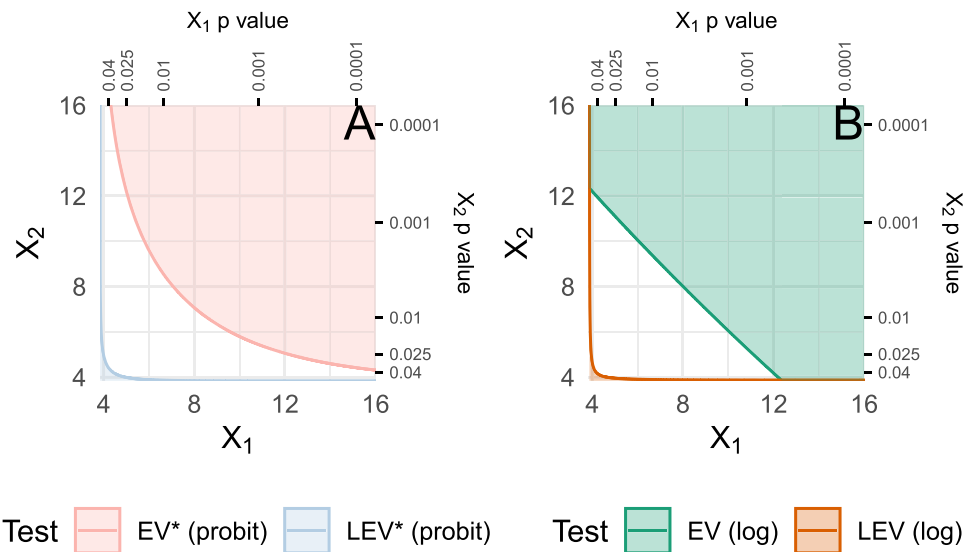


Figure 6. Rejection regions for tests EV^* , LEV^* (panel A), and EV , LEV (panel B) when $k = 2$.

EV^* , something as trivial as rounding to two decimal places can radically alter the result.

4.2.3. General Properties of LEV and LEV^*

For EV and EV^* the null hypothesis was a single point on the edge of the parameter space; for LEV and LEV^* , the null hypothesis that $\lambda_i \geq 2.34, \forall i$ is composite, unbounded, and multiparameter. Lehmann (1952) showed that for such hypotheses, no analytic function can yield rejection bounds for an unbiased test. LEV and LEV^* will thus be biased with alternative regions where the power function is substantially less than the nominal α of the test.

Admissibility with multiparameter tests is difficult to prove and in any case may not yield suitable tests (Perlman and Wu 1999). We are left in a situation where candidate tests are not expected to dominate one another, and tradeoffs between power in one or another region of the parameter space must be considered. Due to the tests' bias, balancing Type I error across the null region and low power in the alternative region must also be considered. The choice of test might be driven by which parts of the parameter space have extremely poor detectability and which do not.

We consider $K = 2$ studies. Figure 6(A) shows the rejection regions for test EV^* and LEV^* . The rejection region for LEV^* is concentrated near small test statistics and along the edges of the space where a single test statistic is small, again showing that a single small test statistic can cause the test to reject. The rejection regions for EV^* and LEV^* are well-separated, indicating that it would be difficult for both tests to reject at the same time (though not impossible; e.g., West et al. 2021).

Figure 6(B) shows the rejection regions for EV and LEV . As with EV^* and LEV^* , the rejection region for LEV is spread along the values just near the significance criterion for each test statistic. The rejection region for EV and LEV overlap much more than for EV^* and LEV^* , meaning that both tests can easily reject at the same time. This is reasonable for tests whose alternatives are “ $\lambda_i > 0$ for some i ” (EV) and “ $\lambda_j < 2.34$ for some j ” (LEV), but is obviously unreasonable for any interpretation in

terms of “[lack of] evidential value” or the “flatness” or “skew” of the distribution of p values.

Figure 7 shows power contours for $K = 2$ studies for tests LEV^* (A) and LEV (B). The figures show the bias of both tests and that the power of one test does not dominate the other. The relative thickness of LEV 's rejection region along the “axes”, as shown in Figure 6(B), translates to higher power when a single test statistic is small. In contrast, the thinness of the tails of the normal transform and hence LEV^* 's rejection region (Figure 6(A)) concentrates more of the power near the origin.

Although by construction, tests LEV and LEV^* are level- α tests of the null hypothesis that “all noncentrality parameters are at least the noncentrality parameter that would yield a one-third probability of statistical significance for the nontruncated test statistic,” we can examine the power contours in Figure 7 and see that there is another hypothesis implied at level α : effectively, “none of the noncentrality parameters are within some generalized distance from the origin.” The generalized distance will depend on the choice of tuning noncentrality being tested against, the transformation used, and the distributions of the test statistics (e.g., χ^2 , F). We also have the aforementioned issue of the bounded parameter space. It is therefore difficult to characterize the inference one would draw from a rejection.

4.3. Power Estimation

Within the official P -curve app, the authors provide a “power estimation” section using same basic logic as the tests EV^* and LEV^* . Their method is not documented in the P -curve papers, but inspection of the code reveals the logic. It is best to regard this not as “power estimation” but rather an estimate of an “overall” noncentrality parameter transformed to a standardized $(0, 1)$ space using a corresponding CDF.

Suppose X_i ($i = 1, \dots, K$), represent K truncated noncentral $\chi^2_1(\lambda, t_\alpha)$ variates. Then

$$\Phi^{-1} \left(\frac{F_\lambda(X_i) - F_\lambda(t_\alpha)}{1 - F_\lambda(t_\alpha)} \right) \sim \text{Normal}(0, 1).$$

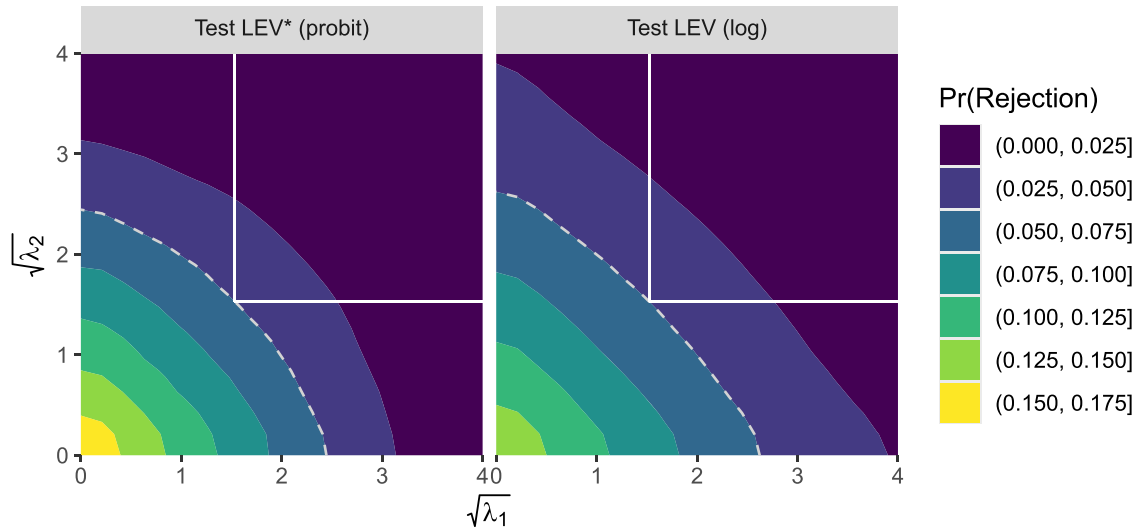


Figure 7. Power contours for tests LEV^* and LEV when $K = 2$. The region in the upper right-hand corner represents the null hypothesis that $\lambda_1, \lambda_2 \geq 2.34$. The dashed line represents the contour where power is equal to the Type I error rate. Axes show $\sqrt{\lambda}$ for clarity.

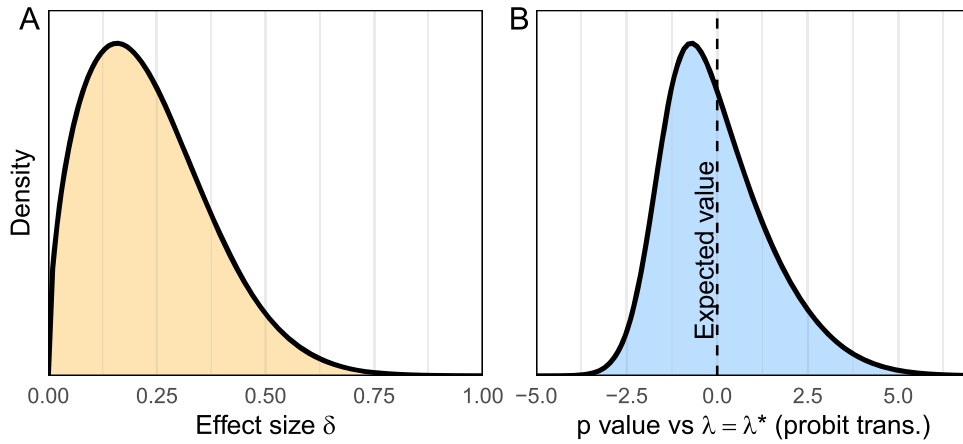


Figure 8. A: Demonstration distribution of effect sizes. B: Resulting distribution of the values summed in the probit power estimation procedure.

where F_λ is the CDF of the noncentral $\chi^2_1(\lambda)$ distribution (nothing in this section depends on this particular assumption; we make it for demonstration). Simonsohn et al. choose a noncentrality parameter λ^* that yields:

$$\sum_{k=1}^K \left[\Phi^{-1} \left(\frac{F_{\lambda^*}(X_i) - F_{\lambda^*}(t_\alpha)}{1 - F_{\lambda^*}(t_\alpha)} \right) \right] = 0.$$

One way to think about λ^* is in terms of test LEV^* : it is the null λ' we would choose for test LEV^* that would yield an LEV^* test statistic of $Z = 0$, a kind of estimate of an overall λ . Then, $q^* = 1 - F_{\lambda^*}(t_\alpha)$ is taken as an estimate of the studies' "power" if they were run without publication bias (and hence without the need for "correction").

Assume there is a population p_λ of noncentrality parameters from which independent sample values $\lambda_1, \dots, \lambda_K$ are drawn. For each λ_i , we then draw a corresponding truncated test statistic X_i . The "average power" of the studies in the population is $E_\lambda [1 - F_\lambda(t_\alpha)]$. We can examine the properties of q^* as an estimate of the population average power. Specifically, we can examine the consistency of q^* as $K \rightarrow \infty$. If the estimator is not consistent, then even arbitrarily large sets of studies may not yield estimates close to the true "average power".

As $K \rightarrow \infty$, $\lambda^* \rightarrow \lambda'$ such that

$$E \left[\Phi^{-1} \left(\frac{F_{\lambda'}(x_i) - F_{\lambda'}(t_\alpha)}{1 - F_{\lambda'}(t_\alpha)} \right) \right] = 0. \quad (1)$$

where the expectation is taken over X and λ . If all $\lambda_i = \lambda$ (no variability) then the estimate will be weakly consistent (proof in the supplement). If we assume variability in the true noncentrality parameters, then the estimate is not generally consistent. Although there will be a λ^* that yields 0 expectation in (1), it needn't yield $E_\lambda [1 - F_\lambda(t_\alpha)] = 1 - F_{\lambda^*}(t_\alpha)$; hence, the estimator is asymptotically biased.

As a demonstration, consider a one-sample design with a Normally distributed test statistic z with mean $\delta\sqrt{N}$ and variance 1, where N is an effective sample size. Assume for all studies, $N = 100$ and assume that variability in the effect size such that $\lambda = \delta^2 N_{eff}$ has an Exponential distribution with a scale of 10 (we could as easily vary the sample size while keeping the effect size constant, or vary both). The corresponding distribution of δ is shown in Figure 8(A). Under these assumptions, the marginal probability of observing a statistically significant effect as $K \rightarrow \infty$ is 0.664.

The λ^* that meets the condition in (1) is $\lambda^* = 9.493$. As can be seen from the density in Figure 8(B), the test statistic

has expectation 0, but is not normally distributed; it has a longer right tail. The method had no chance of arriving at the correct answer: the P -curve only uses the mean of the test statistic distribution, not the whole distribution, implicitly assuming normality. Any situation more complicated than a single λ value will be effectively impossible for the procedure. The estimated power is 0.869 which is 31% larger than it should be. Note that this bias is asymptotic: the estimator is wildly inconsistent. Note also that this estimate depends critically on the probit transformation; other methods (e.g., based on LEV) will yield different “power estimates,” even asymptotically.

In a discussion of the use of a related P -curve method to estimate effect sizes, the P -curve authors (Simmons, Nelson, and Simonsohn 2018) claim that the P -curve can estimate average “power” even when there is heterogeneity in effect sizes. What explains the discrepancy between our demonstration that it fails even in the large sample limit, and their claim that heterogeneity poses no problem for the P -curve? In short, their argument was based only on seven simulations and no formal analysis. One can replicate our demonstration with their own code, suggesting that they simply did not find a simulation violating their intuition. This shows the downside of relying solely on simulation to support a method: it is limited by the simulator’s imagination and confirmation bias. Formal analysis is harder, but can more clearly define a method’s limits.

5. Conclusion

Based on our analysis of the P -curve’s properties, we offer several concrete recommendations. First: do not use Simonsohn, Simmons, and Nelson’s (2015) compound decision rule. This would reduce, but not fully eliminate, problems of nonmonotonicity in the evidence. Second: do not use the test EV^* , because the probit transformation is inappropriate for the random variables used in the test. Third: do not use tests LEV/LEV^* or LS/LS^* due to their extreme sensitivity. Fourth: do not use the “power” estimates from the P -curve, because these estimates are not generally consistent. Fifth: abandon misleading interpretations of the P -curve tests in terms of “skew” and “evidential value”; instead, focus on the null hypotheses that are actually rejected. We are left only with test EV (if properly interpreted). Test EV was eliminated from Simonsohn et al.’s online app, but our online app computes it. We also suggest, however, temporarily avoiding this test until more is understood about admissibility with $F(1, \nu_2)$ statistics and combining across test statistic families.

From a conceptual perspective, it is more challenging to offer concrete recommendations. We recommend that future work in this area focus on directly evaluating the higher-order properties of whole p value distributions, rather than testing a simple transformed average of truncated p values. Re-imaginings should also be based on explicit models of cheating behavior that the developers wish to detect.

Given what is needed to improve the P -curve tests, we do not recommend their use in their current form. Their statistical properties are problematic and it is not clear what substantive conclusions they afford. Given the stated purpose of the P -curve—evaluating the trustworthiness of scientific literatures—the stakes are too high to use tests with such poor, or poorly understood, properties.

Users of the P -curve procedure may object on practical grounds: the tests seem to agree with what they suspect from a histogram of p values. Although the tests are poorly constructed, the results are still driven by patterns in the data, and these patterns overlap with those one might notice in such a histogram. But if the justification of the method cannot rest on formal principles—and we argue the formal justification is shaky at best—and proponents of the method decide instead to justify conclusions via agreement with visual inspection, this raises the question of why the test was necessary in the first place.

As a final point, we suggest that meta-scientists be more skeptical of procedures like the P -curve in the meta-scientific literature. Papers introducing them are often light on statistical exposition, using metaphors and a few simulations to make sweeping arguments. Simulation is a powerful tool and can help build intuition, but it is not a substitute for formal analysis. Simulation may provide hints of problems with a procedure, but only if the simulator’s formal knowledge helps guide the choice of simulations. A simulator might quit after running a few simulations that tell them what they think is true while problems remain uncovered. Given the implications of poor forensic procedures for science, all such procedures demand deeper formal scrutiny.

Supplementary Materials

Supplementary materials contain further description of the literature review, proofs of primary theorems, and further discussion of test LS^* . All code and data are available at https://github.com/richarddmores/MoreyDavisStober_pcurveASA.

Acknowledgments

The authors thank Berna Devezer, Jon Gillard, R. Matthew Montoya, and Christopher K. Wikle for helpful comments about early versions of this article.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

C.P. D-S. was supported by U. S. Department of Defense grant #W81XWH2110173.

References

- Abelson, R. P. (1995), *Statistics As Principled Argument* (1st US ed.), Hillsdale, NJ: Psychology Press. [3,6]
- Berkson, J. (1938), “Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test,” *Journal of the American Statistical Association*, 33, 526–536. [4]
- Birnbaum, A. (1954), “Combining Independent Tests of Significance,” *Journal of the American Statistical Association*, 49, 559–574. [6,8]
- Bishop, D. V., and Thompson, P. A. (2016), “Problems in Using p -curve Analysis and Text-Mining to Detect Rate of p -hacking and Evidential Value,” *PeerJ*, 4, e1715. [1]
- Cadario, R., and Chandon, P. (2020), “Which Healthy Eating Nudges Work Best? A Meta-Analysis of Field Experiments,” *Marketing Science*, 39, 465–486. [1]
- Carney, D. R., Cuddy, A. J., and Yap, A. J. (2010), “Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance,” *Psychological Science*, 21, 1363–1368. [1]

- Cuddy, A. J., Schultz, S. J., and Fosse, N. E. (2018), "P-curving a More Comprehensive Body of Research on Postural Feedback Reveals Clear Evidential Value for Power-Posing Effects: Reply to Simmons and Simonsohn (2017)," *Psychological Science*, 29, 656–666. [1]
- Dominus, S. (2017), "When the Revolution Came for Amy Cuddy," *The New York Times Magazine*. [1]
- Edwards, A. W. F. (1986), "Are Mendel's Results Really Too Close?" *Biological Reviews*, 61, 295–312. [9]
- Erdfelder, E., and Heck, D. W. (2019), "Detecting Evidential Value and p-Hacking With the p-Curve Tool: A Word of Caution," *Zeitschrift für Psychologie*, 227, 249–260. [1]
- Fisher, R. A. (1936), "Has Mendel's Work Been Rediscovered?" *Annals of Science*, 1, 115–137. [9]
- Gelman, A., and O'Rourke, K. (2014), "Discussion: Difficulties in Making Inferences about Scientific Truth from Distributions of Published p-values," *Biostatistics*, 15, 18–23. [2]
- Hosseini-Kamkar, N., Lowe, C., and Morton, J. B. (2021), "The Differential Calibration of the HPA Axis as a Function of Trauma Versus Adversity: A Systematic Review and p-curve Meta-Analyses," *Neuroscience & Biobehavioral Reviews*, 127, 54–135. [1]
- Kozioł, J. A., and Perlman, M. D. (1978), "Combining Independent Chi-Squared Tests," *Journal of the American Statistical Association*, 73, 753–763. [7]
- Lehmann, E. L. (1952), "Testing Multiparameter Hypotheses," *The Annals of Mathematical Statistics*, 23, 541–552. [10]
- Marden, J. I. (1982), "Combining Independent Noncentral Chi Squared or F Tests," *The Annals of Statistics*, 10, 266–277. [6,7,8]
- (1985), "Combining Independent One-Sided Noncentral t or Normal Mean Tests," *The Annals of Statistics*, 13, 1535–1553. [3]
- Matthes, T. K., and Truax, D. R. (1967), "Tests of Composite Hypotheses for the Multivariate Exponential Family," *The Annals of Mathematical Statistics*, 38, 681–697. [6]
- 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, 730–749. [2]
- Montoya, R. M., Kershaw, C., and Jurgens, C. T. (2024), "The Inconsistency of p-curve: Testing its Reliability Using the Power Pose and HPA Debates," *PLoS One*, 19, e0305193. [1]
- Perlman, M. D., and Wu, L. (1999), "The Emperor's New Tests," *Statistical Science*, 14, 355–369. [10]
- Rusz, D., Le Pelley, M. E., Kompier, M. A. J., Mait, L., and Bijleveld, E. (2020), "Reward-Driven Distraction: A Meta-Analysis," *Psychological Bulletin*, 146, 872–899. [4]
- Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366. [1]
- (2018), "P-curve Handles Heterogeneity Just Fine," available at <https://datacolada.org/67>. [12]
- Simmons, J. P., and Simonsohn, U. (2017), "Power Posing: P-Curving the Evidence," *Psychological Science*, 28, 687–693. [1]
- Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014a), "P-curve: A Key to the File-Drawer," *Journal of Experimental Psychology: General*, 143, 534–547. [1,2,3,4,7,9]
- (2014b), "P-curve and Effect Size: Correcting for Publication Bias Using Only Significant Results," *Perspectives on Psychological Science*, 9, 666–681. [2,4]
- (2017), "The p-curve App 4.06," <https://www.p-curve.com/app4/>. [3]
- Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2015), "Better P-curves: Making P-curve Analysis More Robust to Errors, Fraud, and Ambitious P-hacking, a Reply to Ulrich and Miller (2015)," *Journal of Experimental Psychology: General*, 144, 1146–1152. [1,2,3,4,5,6,7,8,12]
- Stuart, M. T., Colaço, D., and Machery, E. (2019), "P-curving x-phi: Does Experimental Philosophy Have Evidential Value?" *Analysis*, 79, 669–684. [1,4]
- Ulrich, R., and Miller, J. (2015), "P-hacking by Post Hoc Selection with Multiple Opportunities: Detectability by Skewness Test?: Comment on Simonsohn, Nelson, and Simmons (2014)," *Journal of Experimental Psychology: General*, 144, 1137–1145. [1]
- (2018), "Some Properties of p-curves, with an Application to Gradual Publication Bias," *Psychological Methods*, 23, 546–560. [2]
- van Assen, M. A. L. M., van Aert, R. C. M., and Wicherts, J. M. (2015), "Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies," *Psychological Methods*, 20, 293–309. [3]
- West, S. J., Hyatt, C. S., Miller, J. D., and Chester, D. S. (2021), "P-Curve Analysis of the Taylor Aggression Paradigm: Estimating Evidentiary Value and Statistical Power across 50 Years of Research," *Aggressive Behavior*, 47, 183–193. [4,10]
- Wilde, V. K. (2022), "Neonatal Jaundice and Autism: Precautionary Principle Invocation Overdue," *Cureus*, 14, e22512. [4]