Journal of Experimental Psychology: General
2015, Vol. 144, No. 6, 1146–1152

0096-3445/15/$12.00 http://dx.doi.org/10.1037/xge0000104

REPLY

Better *P*-Curves: Making *P*-Curve Analysis More Robust To Errors, Fraud, and Ambitious *P*-Hacking, A Reply To Ulrich and Miller (2015)

Uri Simonsohn and Joseph P. Simmons
University of Pennsylvania

Leif D. Nelson
University of California, Berkeley

When studies examine true effects, they generate right-skewed *p*-curves, distributions of statistically significant results with more low (.01 s) than high (.04 s) *p* values. What else can cause a right-skewed *p*-curve? First, we consider the possibility that researchers report only the smallest significant *p* value (as conjectured by Ulrich & Miller, 2015), concluding that it is a very uncommon problem. We then consider more common problems, including (a) *p*-curvers selecting the wrong *p* values, (b) fake data, (c) honest errors, and (d) ambitiously *p*-hacked (beyond $p < .05$) results. We evaluate the impact of these common problems on the validity of *p*-curve analysis, and provide practical solutions that substantially increase its robustness.

Keywords: *p*-curve, publication bias, *p*-hacking

Because statistically significant results are more likely to be published than nonsignificant ones, published scientific evidence is biased. As a consequence, it is difficult to know whether a set of significant findings is indicative of a true effect or whether it reflects nothing more than the selective reporting of significant results. We developed *p*-curve analysis to mitigate this problem (Simonsohn, Nelson, & Simmons, 2014). *P*-Curve is the distribution of statistically significant *p* values (*p*s $< .05$). True effects, those that differ from zero, lead to right-skewed *p*-curves (e.g., more .01s than .04s). Nonexistent effects lead to flat *p*-curves (as many .01s as .04s). And many forms of *p*-hacking, the selective reporting of *analyses* that are $p < .05$, lead to left-skewed *p*-curves (more .04s than .01s).[1] When the observed *p*-curve for a set of studies is right-skewed, we conclude that it contains evidential value; that is, we rule out selective reporting of statistically significant analyses and studies as the only explanation for the results.

Ulrich and Miller (2015) identify a problem with this approach, a problem we had not paid enough attention to until reading their article. The concern is simplest to explain with a stylized example. Imagine a researcher who runs a two-cell experiment investigating the anchoring effect (as in Tversky & Kahneman, 1974). Every participant estimates the length of the Mississippi river after considering either a small or large anchor value. The observed anchoring effect is $p < .05$, whether one controls for participant gender or not, and whether the dependent variable is log-transformed or not. Which of these four significant results will the authors report?

P-Curve analysis assumes that among $p < .05$ results, the magnitude of the *p* value does not affect whether it gets reported. The assumption is met if researchers report all significant results, a random subset, or more likely, the subset that makes for a more compelling or fluent write-up (e.g., the simplest statistical test, or a test used elsewhere in the article).[2]

Ulrich and Miller (2015) correctly note than if instead researchers systematically *only* report the *smallest p* value obtained, then *p*-curve would be right-skewed even in the absence of an effect. Their reasoning is straightforward: Reported *p* values get smaller when only the smallest are reported.

In this article we first respond to Ulrich and Miller's critique, and then we extend it. In our response, we explain why we believe the problem they identify to be quite rare. In our extension, we consider and provide practical solutions to more common problems that may invalidate *p*-curve analyses.

Reporting Only the Smallest *p* Value Is a Rare Form of *P*-Hacking

For at least two reasons, we believe that in the vast majority of cases, researchers are unlikely to report only the smallest *p* value. First, doing so will typically require reporting analyses that are unusual, undermining the credibility of the results. For example,

Uri Simonsohn and Joseph P. Simmons, The Wharton School, University of Pennsylvania; Leif D. Nelson, Haas School of Business, University of California, Berkeley.

Correspondence concerning this article should be addressed to Uri Simonsohn, The Wharton School, University of Pennsylvania, 3730 Walnut Street, 500 Huntsman, Philadelphia, PA 19066. E-mail: uws@wharton.upenn.ed

[1] *P*-Hacking with statistically independent tests also leads to a flat *p*-curves; see Supplement 3 in (Simonsohn et al., 2014).

[2] If researchers are "biased" toward reporting the simplest $p < .05$, the expected *p*-curve is still uniform under the null. This is true because all analyses, simple and complex, have uniform *p*-curves under the null, hence complexity is, ex-ante, uncorrelated with *p* values.

the anchoring article that reports a straightforward comparison of means will seem more credible than one that reports only a result with a log transformation and a gender control. When alternative analyses differ in how interesting, compelling, or valid they are, as we believe they almost always do, reporting only the smallest p value undermines the very goal motivating p-hacking in the first place, the goal of reporting persuasive evidence. The researcher is better served by reporting the most persuasive analysis that is significant rather than the one that is the most significant.

Second, p-hacking occurs when honest researchers face ambiguity about which analyses to run, and convince themselves that those leading to better results are the correct ones (see, e.g., Gelman & Loken, 2014; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011; Vazire, 2015). Motivated reasoning requires ambiguity (Kunda, 1990). The behavior simulated by Ulrich and Miller strikes us as too unambiguously exploitative to be the result of motivated reasoning.

For instance, they simulate collecting 32 different measures, running 32 different tests, and reporting the single smallest p value.[3] We doubt that under most circumstances even the most motivated of thinkers could persuade themselves that such behavior is acceptable. We fool ourselves into overeating ice cream by repeatedly taking the "last" trip to the freezer, not by serving all portions at once on the first trip.

Some Evidence

In our original article, we p-curved a set of experiments published in the *Journal of Personality and Social Psychology* (JPSP) with results reported only with a covariate. We conjectured that researchers would leave out the simpler analysis without a covariate only if it was $p > .05$ (see Figure 3A in Simonsohn et al., 2014). The way we understand p-hacking to happen, therefore, predicts a *left*-skewed p-curve for these studies. Ulrich and Miller, in contrast, write "If several possible covariates are available, [authors] may well try them all and then report the one that produces the most significant result" (p. 1138). This form of "p-hacking" predicts a *right*-skewed p-curve (the main point of their critique). The observed p-curve was significantly *left*-skewed.

Franco, Malhotra, and Simonovits (2014) obtained information on all studies conducted using an online platform funded by the National Science Foundation (NSF), and examined which subset of studies was written up, submitted, and published. Their data provide a rare window into what researchers choose to report: only the lowest p value or all the significant ones? We contacted the authors and asked. They responded that "we are . . . finding strong evidence of the 'select all $p < .05$ results'" (Neil Malhotra, personal email, May 6, 2015). At least for the types of research published in *JPSP*, or conducted using the NSF-funded platform, the evidence suggests that researchers do not p-hack by reporting the smallest p value.

Exceptions

For honest researchers to plausibly report the smallest of many p values computed, we believe the following three conditions must be met: (a) many alternative results for a given study must be equally interesting ex-post, (b) the analyses underlying those many alternative results must be equally justifiable ex-ante, and (c) those analyses must be, by the nature of the data, performed simultaneously.

Ulrich and Miller provide only one example in which the type of selective reporting they are concerned about has been documented: early fMRI psychology research. Vul, Harris, Winkielman, and Pashler (2009) found that about 40% of functional magnetic resonance imaging (fMRI) researchers reported results in their articles only for the voxel *most* correlated with a variable of interest (within a region of interest). This example satisfies all three conditions. Any voxel would be considered equally interesting ex-post, the analysis of any voxel would be equally justifiable ex-ante, and the nature of fMRI testing meant that all analyses were conducted simultaneously. We believe gene studies that examine exploratory correlations with behavior traits, and single-cell recording studies, may also occasionally satisfy all three conditions.

It is valuable to be aware of these exceptions, but important to keep in mind that they are exceptions. For the vast majority of psychological research, theories, previous empirical findings, and readers' intuitions deem some results much more interesting, and some analyses much more credible, than others. In addition, analyses are almost always conducted sequentially. Next we consider more likely paths to invalid p-curves.

Problem 1: *P*-Curvers Not Following Directions

In Simonsohn et al. (2014) we explained how p values should be selected from studies for p-curve analysis to be valid. For example, we wrote,

> Most studies report multiple p values, but not all p values should be included in p-curve. Included p values must meet three criteria: (a) test the hypothesis of interest, (b) have a uniform distribution under the null, and (c) be statistically independent of other p values (p. 540).

To help p-curvers apply these principles, Figure 5 in that article shows which test(s) should be selected from the most common experimental designs in psychology (e.g., two-cell, 2×2 factorial design, etc.).[4]

P-Curvers, however, could disregard these directions, resulting in invalid p-curves that distort readers' understanding of the reviewed literatures and of p-curve analysis. *P*-Curvers selecting the wrong tests is, by far, in our view and experience, the biggest threat to the validity of p-curve analysis.

The most extreme violation consists of selecting *all* p values in an article. One example is by Head, Holman, Lanfear, Kahn, and Jennions (2015), who p-curved all p values published in Open Access journals. The article asks an arguably meaningless question—"What is the evidential value of all tests, whether relevant or

[3] They do also simulate less extreme versions (e.g., collecting only two measures and reporting the lowest p value of those two). But these do not really impact p-curve. For example, the share of $p < .01$ goes up from 20% to about 21%. See their Figures 1–3 where $k = 2$.

[4] Figure 5 is actually a table. Thanks to stringent rules, zealous copy-editing, and our shortcomings as negotiators, it is officially a figure.

irrelevant, whether supportive or unsupportive of the hypotheses of interest?" and provides a statistically invalid answer.[5]

Less extreme but more frequent is the erroneous selection of *p* values associated with the testing of interactions. For studies examining *attenuated interactions* ("the effect is bigger here than there") the *p* value of the interaction goes into *p*-curve (and never the *p* values of the simple effects), while for studies examining *reversing interactions* ("the effect is positive here but negative there") the two simple effects go into *p*-curve (and never the *p* value of the interaction).

The first published *p*-curve analysis we are aware of, which reviewed the impact of ovulatory cycle on women's preferences (Gildersleeve, Haselton, & Fales, 2014), provides an example. From a study predicting an attenuated interaction the *p*-curvers selected the *p* value from the bigger simple effect. The correct *p* value to select (the interaction) is 100 times larger (i.e., much less significant) than the incorrectly selected one. Through email, one of the authors, Haselton, indicated they independently had identified this unintentional error and that to the best of their knowledge it is the only one in their review.[6]

Solution

Avoiding this biggest threat to the validity of *p*-curve analysis is simple: Following the guidelines for selecting results from our original article. Editors and reviewers evaluating *p*-curve articles should begin by evaluating the *P*-Curve Disclosure Table (in which authors summarize which test(s) they selected from each study) and uncompromisingly demanding that authors follow the guidelines that lead to valid *p*-curves. *P*-Curve articles without *P*-Curve Disclosure Tables should not be sent out to reviewers.

Problems 2 and 3: Fake Data and Honest Errors

The human factors behind faking data and honest reporting errors could not be more different, but because the solutions to mitigate their impact on *p*-curve are the same, we discuss both in this section.

Fake Data

Articles with fabricated studies need not become influential to heavily bias our understanding of the literature. A article with fake data, never read by anybody, could nevertheless be influential by distorting a meta-analysis that includes it. This is a plausible scenario. In fact, it happened to us in the first systematic *p*-curve analysis we ever conducted (in June of 2011). We had identified a set of just over 20 articles that met a prespecified study selection rule, including one article by Sanna et al. (2011).

Unlike the rest of the studies in the set, Sanna's *p* values were very low: $p = .000004$, $p = .000009$, and $p = .0012$. Surprised by such strong results with very small samples (*n*s ranging from 15–20 per cell), we took a closer look and identified additional anomalies, most notably that the *SD*s were shockingly similar across conditions. This concern eventually led to a rigorous investigation of misconduct, the retraction of the article, and Sanna's resignation (for more information, see Simonsohn, 2013).

On its own, Sanna et al.'s article probably would not have been too influential, but it would nevertheless heavily bias the meta-analysis of any set of articles that included it. To provide a calibration we ran a simple simulation (R Code available from https://osf.io/mbw5g/) in which we created sets of 25 statistically significant results examining an effect that does not exist (i.e., 25 false-positives that in expectation have a flat *p*-curve). Adding one of Sanna et al.'s studies to the mix (the one with the median *p* value) raises the probability that the overall *p*-curve will be significantly (and falsely) right-skewed from the nominal 5% to 38%. Including all three studies increases the probability of a false right-skew to a heart-stopping 98%. One fake article by Sanna assures virtually any literature of "evidential value."

It would be extraordinarily coincidental for us to have come across the only fraudulent article in psychology in our very first *p*-curve analysis. Indeed, since 2011 we have come across at least five articles (by different authors) with results that seem to us at least as indicative of fraud as Sanna's originally did. Separate from these suspicions, since June of 2011 at least three more cases of fraudulent data have come to light in psychology (Smeesters's, Stapel's, and Förster's cases). This sounds bad, but given how serendipitous the discovery of fraud is, how difficult and costly it is to pursue those suspicions, and how high a threshold exists to conclude data are fabricated, probably most studies with fabricated data go undetected. It is worse than it seems.

Honest Errors

Reporting errors are common. Reviews comparing reported test statistics (e.g., "$t(83) = 2.47$") with reported *p* values (e.g., "$p = .016$") find that a substantial share of published results, about 15%, have errors (Bakker & Wicherts, 2011; Berle & Starcevic, 2007; McGuigan, 1995). On the one hand, the vast majority of these would be inconsequential for *p*-curve analyses, both because they are small (e.g., reporting $p = .013$ as $p < .01$), and because errors in reporting *p* values from test statistics are corrected when using the app (www.p-curve.com/app3). However, on the other hand, more consequential errors are not detectable by comparing *p* values with test statistics, nor are they corrected by the app (e.g., original authors miscoding a variable or accidentally reporting the wrong test result). If the detectable levels of sloppiness are taken as evidence for undetectable levels of sloppiness, there is cause for concern.

Solution. Although it is impossible to fully solve these problems, there are things that can be done. First, our field can work toward reducing fraud and honest errors in the first place. The simplest way to achieve this is through transparency: authors should post their raw data, code, and materials, unless they have a compelling reason not to (Simonsohn, 2013). Journals that do not increase transparency requirements for publications are causally, if not morally, responsible for the continued contamination of the scientific record with fraud and sloppiness.

[5] The authors randomly chose one *p* value from each article, the *p* values were hence statistically independent but they were not uniform under the null nor did they test the hypothesis of interest.

[6] The original prediction by Little, Jones, Burt, and Perrett (2007) was that "women would have stronger preference for symmetry at peak fertility when rating for short-term *than for long-term relationships*" (p. 212; emphasis added). Gildersleeve et al. (2014), therefore, should have selected the interaction effect contrasting the impact of fertility on preference for symmetry in short- versus long-term relationships. That test result is $F(1, 201) = 6.54$, $p = .0113$. Instead they selected the simple effect for short-term relationships. That test result was $t(208) = 3.91$, $p = .000125$.

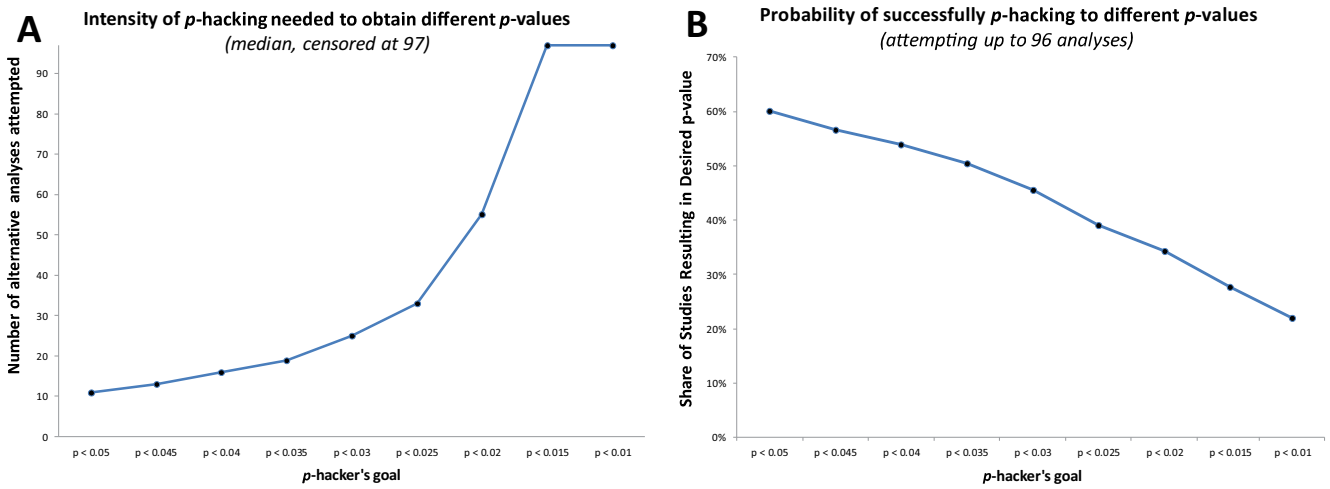

Figure 1. This figure shows the difficulty of *p*-hacking beyond $p < .05$. The results are based on 15,000 simulations that build on those we reported in "False-Positive Psychology" (Simmons et al., 2011), in which authors examine a nonexistent effect and *p*-hack by combining four different analytic decisions (adding 10 observations to 20 already collected, choosing from two dependent variables, dropping one of three conditions, and including a moderating variable). Studies that do not obtain the desired *p* value are assumed to be file-drawered. (R Code available from https://osf.io/mbw5g/). The 96 alternative analyses are assumed to be carried out in random order. See the online article for the color version of this figure.

Second, we can build a more robust *p*-curve. In April of 2015 we released the third version of *p*-curve's online app, www.p-curve.com/app3, making it more robust to extreme results in two important ways. First, we now use Stouffer's method instead of Fisher's method for aggregating results across studies (i.e., combining *pp*-values), because Stouffer's method is less sensitive to a few extreme results (see, e.g., Abelson, 2012, pp. 66–68).[7,8] Earlier we saw that combining one of Sanna's fake studies with 25 false-positive ones lead to a significantly right-skewed *p*-curve 37% of the time. That was using Fisher's method, the approach from our original article. Using Stouffer's method cuts this probability in half (17%).

The online app now also automatically reports what happens when dropping the most extreme *p* values, allowing *p*-curvers and readers to know whether the overall conclusions depend on a few results. See online supplement for details (https://osf.io/mbw5g/).

Third, *p*-curvers should examine the quality of the work they are reviewing, rather than merely copy/pasting results uncritically. When an original article reports problematic results—effects that are too big, statistical results that do not match with the summary statistics, or degrees of freedom that seem erroneously computed, and so forth—*p*-curvers should report *p*-curve with and without those results.

Problem 4: Ambitious *P*-Hacking

What if researchers *p*-hack past $p < .05$? In this section we explore the costs and consequences of such ambitious *p*-hacking and introduce a modification to *p*-curve analysis that makes it robust to plausibly intense levels of it.

The Difficulty of Ambitious *P*-Hacking

In "False-Positive Psychology" (Simmons et al., 2011), we simulated studies conducted by researchers willing to pursue four forms of *p*-hacking (e.g., dropping a condition, dropping a dependent variable, etc.). We found that the probability of obtaining a $p < .05$ result for a nonexistent effect increased from 5% to 61% (see Table 1 in that article).

Relying on the same simulations, Figure 1 shows that the required intensity of *p*-hacking increases exponentially as *p*-hacking gets more ambitious. For instance, once a nonexistent effect has been *p*-hacked to $p < .05$, a researcher would need to attempt nine times as many analyses to achieve $p < .01$. Moreover, as Panel B shows, because there is a limited number of alternative analyses one can do (96 in our simulations), ambitious *p*-hacking often fails.[9]

These results suggest that highly ambitious *p*-hacking (e.g., until reaching $p < .01$) is too difficult to be plausible, especially because there is not a strong incentive to do it. Honest researchers with limited resources and limited motivated reasoning, and without strong incentives to *p*-hack much beyond $p < .05$, usually cannot afford to be that ambitious.

Moderately ambitious *p*-hacking, however, does seem plausible. A $p < .045$ is almost as cheap as a $p < .05$. It is not hard to imagine that honest researchers may continue *p*-hacking past .05 to

[7] *PP*-Values reflect the probability of obtaining at least as extreme a significant *p* value under the null (the *p* value of the *p* value, hence *pp*-value).

[8] For *k* independent *pp*-values, Fisher's test statistic is $\sum -2^{*}\log(pp) \sim \chi^2(2k)$, and Stouffers' is $\sum \Phi^{-1}(pp)/\mathrm{sqrt}(k) \sim N(0,1)$, where Φ^{-1} is the inverse of the *c.d.f.* for the normal distribution (*qnorm*() in R syntax).

[9] We also performed a simulation in which the only form of *p*-hacking consisted of " data peeking," running a *t* test after adding an observation to each of two conditions, starting with $n = 10$ per cell, and ending when the desired *p* value cutoff is obtained or when $n = 100$. Among studies resulting in $p < .05$, getting to $p < .05$ requires 18 peeks, whereas getting to $p < .01$ requires 90 peeks. Overall, a $p < .05$ is obtained 28% of the time, a $p < .01$ only 9%.

avoid, say, an eyebrow-raising $p = .048$. Only past $p < .03$ does *p*-hacking get very expensive very quickly.

Consequences of Moderately Ambitious *P*-Hacking

For an initial calibration, we selected sets of 20 *p* values obtained from the simulations of ambitious *p*-hacking used to generate Figure 1 and submitted them to *p*-curve analysis. In the absence of *p*-hacking, a nonexistent effect has, by definition, a 5% chance of a significant right skew. As researchers *p*-hack trying to get $p < .05$, that probability *drops*, to just 2 in 10,000 in our simulations, because with *p*-hacking, *p*-curve is left-skewed in expectation, so obtaining a significant right-skew becomes less likely.

Ambitious *p*-hacking results in either dropping the higher *p* values, or replacing them with lower ones, increasing the odds of a significant right-skew. When *p*-hacking to $p < .045$, this effect is minimal, increasing the probability of a significant right-skew to about .5% (half a percent). *P*-Hacking to $p < .04$ increases the chances of an erroneous right-skew to a still acceptable 6%. However, it is bad news beyond this point. *P*-Hacking to $p < .035$ raises the right-skew probability to 30%, and *p*-hacking to $p < .03$ raises it to 79%. *P*-Curve confuses *plausible* levels of ambitious *p*-hacking for evidential value unacceptably often.

Solution to Ambitious *P*-Hacking

To deal with moderately ambitious *p*-hacking we focus on the "half *p*-curve," the distribution of $p < .025$ results, asking whether the distribution of *p* values between 0 and .025 is right-skewed. On the one hand, because half *p*-curve does not include barely significant results, it has a lower probability of mistaking ambitious *p*-hacking for evidential value. On the other hand, dropping observations makes the half *p*-curve less powerful, so it has a higher chance of failing to recognize actual evidential value.

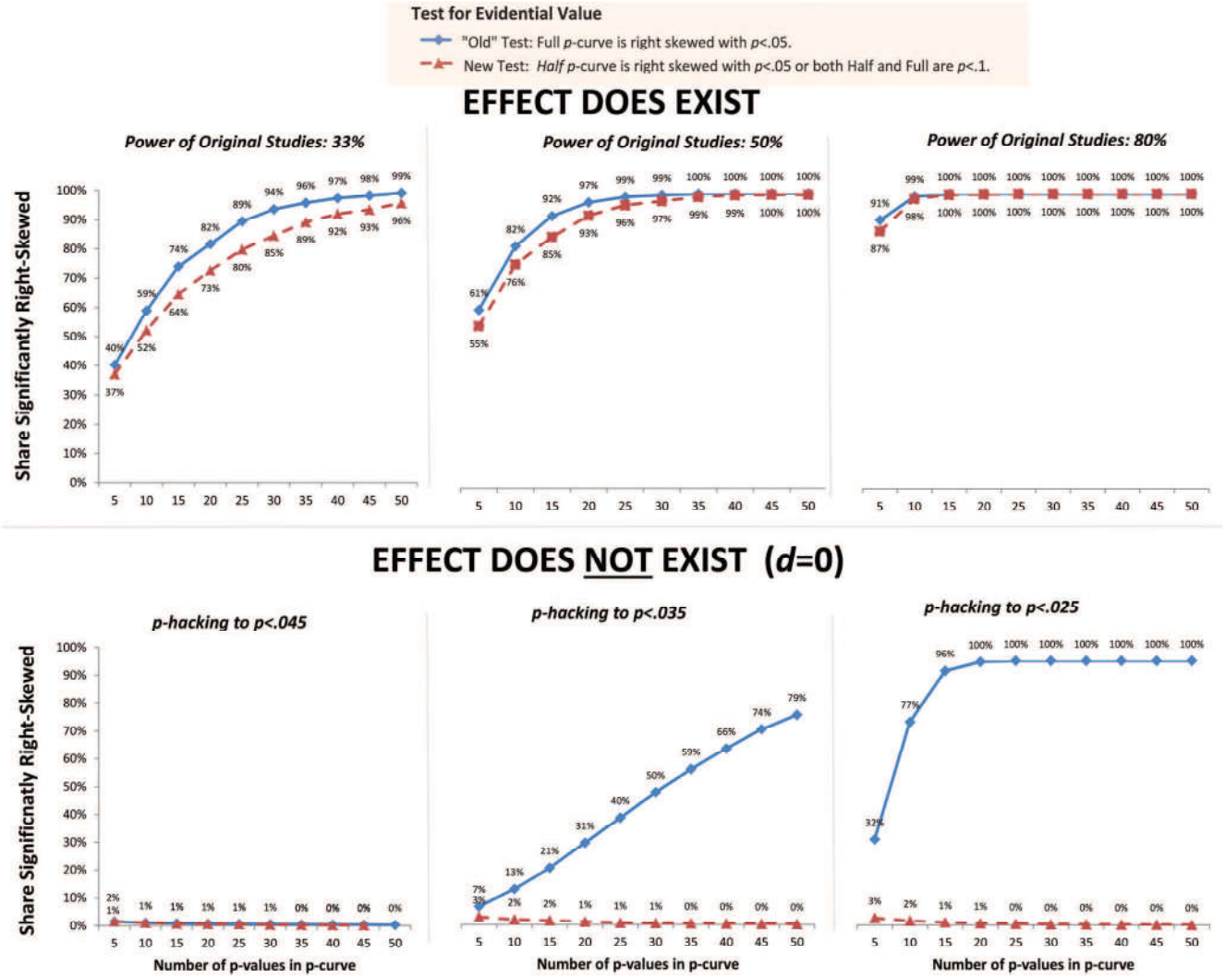

Figure 2. Each plot depicts the share of simulations for which one concludes the underlying studies have evidential value using either the full *p*-curve test (as in Simonsohn et al., 2014) or the novel half/full combination test. Ambitiously *p*-hacked results analyzed via *p*-curve were obtained from the simulations behind Figure 1. (R Code available from https://osf.io/mbw5g/). See the online article for the color version of this figure.

Table 1
Summary of Common Problems That Invalidate P-Curve and Their Solutions

Problem	Solutions
(1) *P*-Curvers select incorrect tests into *p*-curve (e.g., selecting from a study predicting an attenuated interaction the *p*-value of the bigger simple effect, instead of that of the interaction).	*P-Curvers:* Use Figure 5 in Simonsohn, Nelson, and Simmons (2014) to determine which tests to select from each study included in *p*-curve. *Reviewers*: Compare the selection of *p*-values summarized in the submitted "*P*-Curve Disclosure Table" to Figure 5 *Editors*: Before sending articles that include *p*-curve analyses to reviewers, ensure the manuscript includes a "*P*-Curve Disclosure Table."
(2) Original data are fake.	*Journals*: Require the posting of raw data, original materials, and code to analyze the data. *P-Curve app (1):* Aggregate studies using Stouffer's instead of Fisher's method to be less sensitive to few extreme values.
(3) Original results have honest reporting errors.	*p-Curve app* (2): Report how results change if most extreme observations are dropped. *P-Curvers:* Evaluate quality of original work, if errors or improbable results are apparent, report *p*-curve results with and without including those studies.
(4) Ambitious *p*-hacking (e.g., *p*-hacking till $p < .035$).	*New definition of evidential value:* "A set of studies is said to contain evidential value if either the half *p*-curve has a $p < .05$ right-skew test, or both the full and half p-curves have $p < .1$ right-skew tests."

Fortunately, by *combining* the full and half *p*-curves into a single analysis, it is possible to eliminate these false-positive conclusions of evidential value without much of a decrease in power. 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 $p < .05$ right-skew test, *or* both the full and half *p*-curves have $p < .1$ right-skew tests.[10]

In Figure 2 we compare the performance of this new combination test with that of the full *p*-curve alone (the "old" test). The top three panels show that both tests are similarly powered to detect true effects. Only when original research is underpowered at 33% is the difference noticeable, and even then it seems acceptable. With just 5 *p* values *p*-curve has more power than the underlying studies do. The bottom panels show that moderately ambitious *p*-hacking fully invalidates the "old" test (in line with the calibration from above), but the new test is unaffected by it.[11]

Summary

The validity of *p*-curve is threatened by actions original researchers and *p*-curvers can take. The practical solutions put forward in this article, and summarized in Table 1, make *p*-curve vastly more robust to such actions.

[10] As with all cutoffs, it only makes sense to use these as points of reference. A half *p*-curve with $p = .051$ is nearly as good as with $p = .049$, and both tests with $p < .001$ is much stronger than both tests with $p = .099$.

[11] When the true effect is zero and researchers do not *p*-hack (an unlikely combination), the probability that the new test leads to concluding the studies contain evidential value is 6.2% instead of the nominal 5% (R Code: https://osf.io/mbw5g/).

References

Abelson, R. P. (2012). *Statistics as principled argument*. New York, NY: Psychology Press.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43,* 666–678. http://dx.doi.org/10.3758/s13428-011-0089-5

Berle, D., & Starcevic, V. (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research, 16,* 202–207. http://dx.doi.org/10.1002/mpr.225

Franco, A., Malhotra, N., & Simonovits, G. (2014). Social science. Publication bias in the social sciences: Unlocking the file drawer. *Science, 345,* 1502–1505. http://dx.doi.org/10.1126/science.1255484

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102,* 460. http://dx.doi.org/10.1511/2014.111.460

Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014). Meta-analyses and p-curves support robust cycle shifts in women's mate preferences: Reply to Wood and Carden (2014) and Harris, Pashler, and Mickes (2014). *Psychological Bulletin, 140,* 1272–1280. http://dx.doi.org/10.1037/a0037714

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13,* e1002106. http://dx.doi.org/10.1371/journal.pbio.1002106

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524–532. http://dx.doi.org/10.1177/0956797611430953

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480–498. http://dx.doi.org/10.1037/0033-2909.108.3.480

Little, A. C., Jones, B. C., Burt, D. M., & Perrett, D. I. (2007). Preferences for symmetry in faces change across the menstrual cycle. *Biological Psychology, 76,* 209–216. http://dx.doi.org/10.1016/j.biopsycho.2007.08.003

McGuigan, S. M. (1995). The use of statistics in the British Journal of Psychiatry. *The British Journal of Psychiatry, 167,* 683–688. http://dx.doi.org/10.1192/bjp.167.5.683

Sanna, L. J., Chang, E. C., Miceli, P. M., & Lundberg, K. B. (2011). RETRACTED: Rising up to higher virtues: Experiencing elevated physical height uplifts prosocial actions. *Journal of Experimental Social Psychology, 47,* 472–476. http://dx.doi.org/10.1016/j.jesp.2010.12.013

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24,* 1875–1888. http://dx.doi.org/10.1177/0956797613480366

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143,* 534–547. http://dx.doi.org/10.1037/a0033242

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty-heuristics and biases. *Science, 185,* 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124

Ulrich, R., & 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. http://dx.doi.org/10.1037/xge0000086

Vazire, S. (2015). *This is what p-hacking looks like*. Retrieved from http://web.archive.org/web/20150406031852/sometimesimwrong.typepad.com/wrong/2015/02/this-is-what-p-hacking-looks-like.html

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4,* 274–290. http://dx.doi.org/10.1111/j.1745-6924.2009.01125.x

Received June 5, 2015
Revision received July 15, 2015
Accepted July 16, 2015 ■