This draft: 2014 01 15
Forthcoming, Journal of Experimental Psychology: General

P-curve: A Key To The File Drawer

Uri Simonsohn
University of Pennsylvania
The Wharton School
uws@wharton.upenn.edu

Leif D. Nelson
UC Berkeley
Haas School of Business
Leif_nelson@haas.berkeley,edu

Joseph P. Simmons
University of Pennsylvania
The Wharton School
jsimmo@wharton.upenn.edu

Abstract:
Because scientists tend to report only studies (publication bias) or analyses (*p*-hacking) that "work," readers must ask, "Are these effects true, or do they merely reflect selective reporting?" We introduce *p*-curve as a way to answer this question. *P*-curve is the distribution of statistically significant *p*-values for a set of studies (*p*s < .05). Because only true effects are expected to generate right-skewed *p*-curves – containing more low (.01s) than high (.04s) significant *p*-values – only right-skewed *p*-curves are diagnostic of evidential value. By telling us whether we can rule out selective reporting as the sole explanation for a set of findings, *p*-curve offers a solution to the age-old inferential problems caused by file-drawers of failed studies and analyses.

Please: If you see an error in the paper or supplementary materials, contact us directly first.

Supplemental materials, online app, and a *p*-curve user guide available at www.p-curve.com

We would like to thank all of the following people for their thoughtful and helpful comments: Max Bazerman, Jesse Chandler, Noriko Coburn, Geoff Cumming, Clintin Davis-Stober, Ap Dijksterhuis, Ellen Evers, Klaus Fiedler, Geoff Goodwin, Tony Greenwald, Christine Harris, Yoel Inbar, Karim Kassam, Barb Mellers, Michael B. Miller (Minnesota), Julia Minson, Alice Moon, Nathan Novemsky, Hal Pashler, Frank L. Schmidt (Iowa), Phil Tetlock, Wendy Wood, Job Van Wolferen, and Joachim Vosgerau. All remaining errors are our fault.

Scientists tend to publish studies that “work” and to file-drawer those that do not (Rosenthal, 1979). As a result, published evidence is unrepresentative of reality (Ioannidis, 2008; Pashler & Harris, 2012). This is especially problematic when researchers investigate nonexistent effects, as journals tend to publish only the subset of evidence falsely supporting their existence. The publication of false-positives is destructive, leading researchers, policy makers, and funding agencies down false avenues, stifling and potentially reversing scientific progress. Scientists have been grappling with the problem of publication bias for at least 50 years (Sterling, 1959).

One popular intuition is that we should trust a result if it is supported by many different studies. For example, John Ioannidis, an expert on the perils of publication bias, has proposed that “when a pattern is seen repeatedly in a field, the association is probably real, even if its exact extent can be debated” (Ioannidis, 2008, p. 640). This intuition relies on the premise that false-positive findings tend to require many failed attempts. For example, with a significance level of .05 (our assumed threshold for statistical significance throughout this paper), a researcher studying a nonexistent effect will, on average, observe a false-positive only once in 20 studies. Because a *repeatedly* obtained false-positive finding would require implausibly large file-drawers of failed attempts, it would seem that one could be confident that findings with multi-study support are in fact true.

As seductive as this argument seems, its logic breaks down if scientists exploit ambiguity in order to obtain statistically significant results (Cole, 1957; Simmons, Nelson, & Simonsohn, 2011). While collecting and analyzing data, researchers have many decisions to make, including whether to collect more data, which outliers to exclude, which measure(s) to analyze, which covariates to use, etc. If these decisions are not made in advance but rather as the data are being analyzed, then researchers may make them in ways that self-servingly increase their odds of publishing (Kunda, 1990). Thus, rather than file-drawering entire studies, researchers may file-drawer merely the subsets of analyses that produce non-significant results. We refer to such behavior as *p-hacking*.[1]

The practice of *p*-hacking upends assumptions about the number of failed studies required to produce a false-positive finding. Even seemingly conservative levels of *p*-hacking make it easy for researchers to find statistically significant support for nonexistent effects. Indeed, *p*-hacking can allow researchers to get *most* studies to reveal significant relationships between truly unrelated variables (Simmons, et al., 2011).

Thus, although researchers' file drawers may contain relatively few failed and discarded whole studies, they may contain many failed and discarded analyses. It follows that researchers can *repeatedly* obtain evidence supporting a false hypothesis without having an implausible number of failed studies. *P*-hacking, then, invalidates "fail-safe" calculations often used as reassurance against the file-drawer problem (Rosenthal, 1979).

The practices of *p*-hacking and file-drawering mean that a statistically significant finding may reflect selective reporting rather than a true effect. In this paper, we introduce *p*-curve as a way to distinguish between selective reporting and truth. *P*-curve is the distribution of statistically significant *p*-values for a set of independent findings. Its shape is diagnostic of the *evidential value* of that set of findings.[2] We say that a set of significant findings contains evidential value when we can rule out selective reporting as the sole explanation of those findings.

As detailed below, only right-skewed *p*-curves, those with more low (e.g., .01s) than high (e.g., .04s) significant *p*-values, are diagnostic of evidential value. *P*-curves that are not right-

[1] Trying multiple analyses to obtain statistical significance has received many names, including: *bias* (Ioannidis, 2005), *significance chasing* (Ioannidis & Trikalinos, 2007), *data snooping* (White, 2003), *fiddling* (Bross, 1971), *publication bias in situ* (Phillips, 2004) and *specification searching* (Gerber & Malhotra, 2008a). The lack of agreement in terminology may stem from the imprecision of these terms; all of these could mean things quite different from, or capture only a subset of, what the term "*p*-hacking" encapsulates.

[2] In an ongoing project we also show how *p*-curve can be used for effect-size estimation (Nelson, Simonsohn, & Simmons, 2013).

skewed suggest that the set of findings lacks evidential value, and *p*-curves that are left-skewed suggest the presence of intense *p*-hacking.

For inferences from *p*-curve to be valid, studies and *p*-values must be appropriately selected. As outlined in a later section of this article, selected *p*-values must be (1) associated with the hypothesis of interest, (2) statistically independent from other selected *p*-values, and (3) distributed uniform under the null.

Interpreting Evidential Value and Lack Thereof

P-curve's ability to diagnose evidential value represents a critical function: to help distinguish between sets of significant findings that are likely vs. unlikely to be the result of selective reporting.

The inferences one ought to draw from a set of findings containing evidential value are analogous to the inferences one ought to draw from an individual finding being statistically significant. For example, saying that an individual finding is not statistically significant is not the same as saying that the *theory* that predicted a significant finding is wrong. A researcher may fail to obtain statistical significance even when testing a correct theory, because the manipulations were too weak, the measures too noisy, the samples too small, or for other reasons. Similarly, when one concludes from *p*-curve that a set of studies lacks evidential value, one is not saying that the theories predicting supportive evidence are wrong; *p*-curve assesses the reported data, not the theories they are meant to be testing. Thus, both significance and evidential value pertain to the specific operationalizations and samples used in specific studies, not the general theories that those studies test.

Just as an individual finding may be statistically significant even if the theory it tests is incorrect – because the study is flawed (e.g., due to confounds, demand effects, etc.) – a set of studies investigating incorrect theories may nevertheless contain evidential value precisely because that set of studies is flawed. Thus, just like inferences from statistical significance are theoretically important only when one is studying effects that are theoretically important, inferences from *p*-curve are theoretically important only when one *p*-curves a set of findings that are theoretically important.

Finally, just like statistical significance, evidential value does not imply practical significance. One may conclude that a set of studies contains evidential value even though the observed findings are small enough to be negligible (Cohen, 1994), or useless due to failures of internal or external validity (Campbell & Stanley, 1966).

The only objective of significance testing is to rule out chance as a likely explanation for an observed effect. The only objective of testing for evidential value is to rule out selective reporting as a likely explanation for a set of statistically significant findings.

P-curve's many uses

Because many scientific decisions would be helped by knowing whether a set of significant findings contains evidential value, *p*-curve is likely to prove helpful in making many scientific decisions. Readers may use *p*-curve to decide which articles or literatures to read attentively, or as a way to assess which set of contradictory findings is more likely to be correct. Researchers may use *p*-curve to decide which literatures to build on, or which studies to attempt costly replications of. Reviewers may use *p*-curve to decide whether to ask authors to attempt a direct replication prior to accepting a manuscript for publication. Authors may use *p*-curve to explain inconsistencies between their own findings and previously published findings. Policy makers may use *p*-curve to compare the evidential value of studies published under different editorial policies. Indeed, *p*-curve will be useful to anyone who finds it useful to know whether a given set of significant findings is merely the result of selective reporting.[3]

P-curve's shape

In this section we provide an intuitive treatment of the relationship between the evidential value of data and *p*-curve's shape. Supplement 1 presents mathematical derivations based on noncentral distributions of test statistics.

As illustrated below, one can infer whether a set of findings contains evidential value by examining *p*-value distributions. Because *p*-values greater than .05 are often file-drawered and hence unpublished, the published record does not allow one to observe an unbiased sample of the *full* distribution of *p*-values. However, all *p*-values below .05 are potentially publishable and hence observable. This allows us to make unbiased inferences from the distribution of *p*-values that fall below .05, i.e., from *p*-curve.

When a studied effect is nonexistent (i.e., the null hypothesis is true), the expected distribution of *p*-values of independent (continuous) tests is uniform, by definition. A *p*-value indicates how likely one is to observe an outcome at least as extreme as the one observed if the studied effect were nonexistent. Thus, when an effect is nonexistent, $p<.05$ will occur 5% of the time, $p<.04$ will occur 4% of the time, etc. It follows that $.04<p<.05$ should occur 1% of the time, $.03<p<.04$ should occur 1% of the time, etc. When a studied effect is nonexistent, every *p*-value is equally likely to be observed, and *p*-curve will be uniform (Figure 1A).[4]

[3] Because *p*-curve is solely derived from statistically significant findings, it cannot be used to assess the validity of a set of non-significant findings.

[4] Figure 1 assumes normally distributed data. As shown in Supplement 2, *p*-curve's expected shape is robust to other distributional assumptions.

When a studied effect does exist (i.e., the null is false), the expected distribution of *p*-values of independent tests is right-skewed, i.e., more *p*s<.025 than .025<*p*s<.05 (Cumming, 2008; Hung et al., 1997; Lehman, 1986; Wallis, 1942). To get an intuition for this, imagine a researcher studying the effect of gender on height with a sample size of 100,000. Because men are so reliably taller than women, the researcher would be more likely to find strongly significant evidence for this effect ($p < .01$) than to find weakly significant evidence for this effect ($.04 < p < .05$). Investigations of smaller effects with smaller samples are simply less extreme versions of this scenario. Their expected *p*-curves will fall between this extreme example and the uniform for a null effect and so will also be right-skewed. Figures 1B-1D show that when an effect exists, no matter whether it is large or small, *p*-curve is right-skewed.[5]

Like statistical power, *p*-curve's shape is a function only of effect size and sample size (see Supplement 1). Moreover, as Figures 1B-1D show, expected *p*-curves become more markedly right-skewed as true power increases. For example, they show that studies truly powered at 46% ($d = .6$, $n = 20$) have about six $p<.01$ for every $.04<p<.05$, whereas studies truly powered at 79% ($d = .9$, $n = 20$) have about eighteen $p<.01$ for every $.04<p<.05$.[6]

If a researcher *p*-hacks, attempting additional analyses in order to turn a nonsignificant result into a significant one, then *p*-curve's expected shape changes. When *p*-hacking, researchers are unlikely to pursue the lowest possible *p*-value; rather, we suspect that *p*-hacking frequently stops upon obtaining significance. Accordingly, a disproportionate share of *p*-hacked *p*-values will be higher rather than lower.[7]

[5] Wallis (1942, pp. 237-238) provides the most intuitive treatment we have found for right-skewed *p*-curves.

[6] Discussions of statistical power typically involve arbitrary effect sizes (Cohen, 1962, 1992) or those estimated – with publication bias – from a given sample (Ioannidis & Trikalinos, 2007). Throughout this article we talk about *true* power, the likelihood of obtaining statistical significance given the *true* magnitude of the effect size being studied.

[7] Supplement 3 provides a formal treatment of the impact of *p*-hacking on *p*-curve.

For example, consider a researcher who *p*-hacks by analyzing data every 5 per-condition participants and ceases upon obtaining significance. Figure 1E shows that when the null hypothesis is true ($d = 0$), the expected *p*-curve is actually left-skewed.[8]

Because investigating true effects does not guarantee a statistically significant result, researchers may *p*-hack to find statistically significant evidence for true effects as well. This is especially likely when researchers conduct underpowered studies, those with a low probability of obtaining statistically significant evidence for an effect that does in fact exist. In this case *p*-curve will combine a right-skewed curve (from the true effect) with a left-skewed one (from *p*-hacking). The shape of the resulting *p*-curve will depend on how much power the study had (before *p*-hacking) and the intensity of *p*-hacking.

A set of studies with sufficient-enough power and mild-enough *p*-hacking will still tend to produce a right-skewed *p*-curve, while a set of studies with low-enough power and intense-enough *p*-hacking will tend to produce a left-skewed *p*-curve. Figures 1F-1H show the results of simulations demonstrating that the simulated intensity of *p*-hacking produces a markedly right-skewed curve for a set of studies that are appropriately powered (79%), a moderately right-skewed curve for a set of studies that are underpowered (45%), and a barely left-skewed curve for a set of studies that are drastically underpowered (14%).

In sum, the reality of a set of effects being studied interacts with researcher behavior to shape the expected *p*-curve in the following way:

- Sets of studies investigating effects that exist are expected to produce right-skewed *p*-curves.
- Sets of studies investigating effects that do not exist are expected to produce uniform *p*-curves.

[8] As we show in Supplement 3, the effect of *p*-hacking on *p*-curve depends on one key parameter: the expected correlation between *p*-values of consecutive analyses. Thus, although our simulations focus only on *p*-hacking via data peeking, these results generalize to other forms of *p*-hacking.

- Sets of studies that are intensely *p*-hacked are expected to produce left-skewed *p*-curves.

Heterogeneity and *p*-curve

Because *p*-curve is expected to be right-skewed for any non-zero effect size, no matter how researchers choose their sample sizes, all combinations of studies for which at least some effects exist are expected to produce right-skewed *p*-curves. This is true regardless of the degree of variability in sample sizes or effect sizes in the set. Figure 2 reports expected *p*-curves for a set of 10 heterogeneous studies: No matter whether sample size and effect size are negatively, positively, or uncorrelated across studies, *p*-curve is expected to be right-skewed.

*** Figure 2 ***

When a set of studies has a right-skewed *p*-curve, we infer that the set has evidential value, which means that at least some studies in the set have evidential value. Inference from *p*-curve is hence once again analogous to inference about statistical significance for a single finding. When the *average* of the dependent variable significantly differs across conditions, it means that at least *some* observations were influenced by the manipulation. A right-skewed *p*-curve does not imply all studies have evidential value, just as a significant difference across conditions does not imply that all observations were influenced by the manipulation.

If a set of studies can be meaningfully partitioned into subsets, it is the job of the individual who is *p*-curving to determine if such partitioning should be performed, in much the same way that it is the job of the person analyzing experimental results to decide if a given effect should be tested on all observations combined, or if a moderating factor is worth exploring. Heterogeneity, then, poses a challenge of interpretation, not of statistical inference. *P*-curve answers whether selective reporting can be ruled out as an explanation for a set of significant findings; a person decides which set of significant findings to submit to a *p*-curve test.

Statistical inference with *p*-curve

This section describes how to assess if the observed shape of *p*-curve is statistically significant. A web-based application that conducts all necessary calculations is available at http://p-curve.com; users need only input the set of statistical results for the studies of interest and all calculations are performed automatically.

Readers not interested in the statistical details of *p*-curve may skip or skim this section.

Does a set of studies contain evidential value? Testing for right-skew

This section describes how to test if *p*-curve is significantly right-skewed; all calculations are analogous for testing if *p*-curve is significantly left-skewed. A simple method consists of dichotomizing *p*-values as *high* (p>.025) versus *low* (p<.025), and submitting the uniform null (50% high) to a binomial test. For example, if five out of five findings were p<.025, *p*-curve would be significantly right-skewed (p=$.5^5$=.03125). This test has the benefits of being simple and also resistant to extreme *p*-values (e.g., to a study with p=.04999).

This test, however, ignores variation in *p*-values within the *high* and *low* bins and is hence inefficient. To account for this variation, we propose a second method that treats *p*-values as test statistics themselves, and we focus on this method throughout. This method has two steps. First, one computes, for each significant *p*-value, the probability of observing a significant *p*-value at least as extreme if the null were true. This is the *p*-value of the *p*-value, and so we refer to it as the *pp-value*. For continuous test statistics (e.g., *t*-tests), the *pp*-value for the null of a uniform *p*-curve against a right-skew alternative is simply $pp=\frac{p}{.05}$. For example, under the null that *p*-curve is uniform there is a 20% chance that a significant *p*-value will be p<.01; thus, p=.01 corresponds to a *pp*-value of .2. Similarly, there is a 40% chance that p<.02, and so p=.02 corresponds to a *pp*-value of .40.

The second step is to aggregate the *pp*-values, which we do using Fisher's method.[9] This yields an overall χ^2 test for skew, with twice as many degrees of freedom as there are *p*-values. For example, imagine a set of three studies, with *p*-values of .001, .002, and .04, and thus *pp*-values of .02, .04 and .8. Fisher's method generates an aggregate $\chi^2(6)=14.71$, $p=.022$, a significantly right-skewed *p*-curve.[10]

Does a set of studies lack evidential value? Testing for power of 33%

Logic underlying the test. What should one conclude when a *p*-curve is not significantly right-skewed? In general, null findings may result from either the precise estimate of a very small or nonexistent effect, or from the noisy estimate of an effect that may be small or large. In *p*-curve's case, a null finding (a *p*-curve that is not significantly right-skewed) may indicate that the set of studies lacks evidential value, or that there is not enough information (i.e., not enough *p*-values) to make inferences about evidential value.

In general, one can distinguish between these two alternative accounts of a null finding by considering a different null hypothesis: not that the effect is zero, but that it is *very small* instead (see e.g., Cohen, 1988, pp. 16-17; Greenwald, 1975; Hodges & Lehmann, 1954; Serlin & Lapsley, 1985). Rejecting a very small effect, against the alternative of an *even smaller* one, allows one to go beyond merely saying that an effect is not significantly different from zero; it allows one to conclude that an effect is not larger than negligible.

When *p*-curve is not significantly right-skewed, we propose testing whether it is flatter than

[9] Fisher's method capitalizes on the fact that -2 times the sum of the natural log of each of *k* uniform distributions (i.e., *p*-values under the null) is distributed $\chi^2(2k)$.

[10] The *ln* of .02, .04 and .8 are -3.91, -3.21, and -.22, respectively. Their sum is -7.35, and thus the overall test becomes $\chi^2(6)=14.71$.

one would expect if studies were powered at 33%.[11] Studies with such extremely low power fail 2 out of 3 times, and so, of course, do direct replications that use identical sample sizes. If a set of studies has a *p*-curve that is significantly flatter than even that produced by studies that are powered at 33%, then we interpret the set of findings as lacking evidential value. Essentially, we would be concluding that the effects of interest, even if they exist, are too small for existing samples, and that researchers interested in the phenomenon would need to conduct new studies with better powered samples to allow for data with evidential value to start accumulating. If the null of such a small effect is *not* rejected, then *p*-curve is inconclusive. In this case, more *p*-values are needed to determine whether or not the studies contain evidential value.

Implementation of the test. Recall that expected *p*-curves are, like statistical power, a function of only sample and effect size. Because of this, it is straightforward to test the null hypothesis of a very small effect: The procedure is the same as for testing right-skew, except that *pp*-values are recomputed for the expected *p*-curves given a power of 33% and the study's sample size. (As we detail in Supplement 1, this is accomplished by relying on noncentral distributions).

For example, we know that t-tests with 38 degrees of freedom and powered at 33% will result in 57.6% of significant *p*-values being greater than .0l, and so the *pp*-value corresponding to *p*=.01 for a *t*(38) test would be *pp*=.576. Similarly, only 10.4% of significant *p*-values are expected to be greater than .04, and so the *pp*-value corresponding to *p*=.04 for this test would be *pp*=.104.[12]

[11] As is the case for all cutoffs proposed for statistical analyses (e.g., "*p*<.05 is significant", "power of 80% is appropriate", "d>.8 is a large effect", "a Bayes factor <3 indicates anecdotal evidence"), the line we draw at power of one-third is merely a suggestion. Readers may choose a different cutoff point for their analyses. The principle that one may classify data as lacking evidential value when the studies are too underpowered is more important than exactly where we draw the line for concluding that effects are "too underpowered."

[12] In this case, we are interested in whether the observed *p*-values are too high, so we compute *pp*-values for observing a *larger p*-value.

Consider three studies, each with 20 subjects per cell, producing t-tests with *p*-values of .040, .045, and .049.[13] For the right-skew test, the *pp*-values would be .8, .9 and .98, respectively, leading to an overall $\chi^2(6)=0.97$, *p*=.99; in this case, *p*-curve is far from significantly right-skewed. We then proceed to testing the null of a very small effect. Under the null of 33% power the *pp*-values are .104, .0502 and .0098, respectively, leading to an overall $\chi^2(6)=19.76$, *p*=.003; in this case, we reject the null of a small effect against the alternative of an even smaller one. Thus, we conclude that this set of studies lacks evidential value; either the studied effects do not exist, or they are too small for us to rule out selective reporting as an alternative explanation.

If, instead, the three *p*-values had been .01, .01 and .02, then *p*-curve would again not be significantly right-skewed, $\chi^2(6)=8.27$, *p*=.22. However, in this case, we could not reject a very small effect, $\chi^2(6)=4.16$, *p*=.65. Our conclusion would be that *p*-curve is too noisy to allow for inferences from *those* three *p*-values.

A Demonstration

We demonstrate the use of *p*-curve by analyzing two sets of diverse findings published in the *Journal of Personality and Social Psychology (JPSP).* We hypothesized that one set was likely to have been *p*-hacked (and thus less likely to contain evidential value), and that the other set was unlikely to have been. We used *p*-curve to test these two hypotheses.

The first set consisted of experiments in which, despite participants having been randomly assigned to conditions, the authors reported results *only* with a covariate. To generate this set, we searched *JPSP*'s archives for the words "experiment" and "covariate," and applied a set of pre-determined selection rules until we located 20 studies with significant *p*-values (see Supplement 5 for details). The decision to collect *p*-values from 20 studies was made in advance.

[13] None of the analyses in this paper require the sample sizes to be the same across studies. We have made them the same only to simplify our example.

The second set consisted of 20 studies whose main text gave no hints of *p*-hacking. To generate this set we searched *JPSP*'s archives for experiments described without terms possibly associated with *p*-hacking (e.g., "excluded", "covariate", and "transform"; see Supplement 5).

It is worth emphasizing that there is nothing wrong with controlling for a covariate, even in a study with random assignment. In fact, there is plenty right with it. Covariates may reduce noise, enhancing the statistical power to detect an effect. We were suspicious of experiments reporting an effect only with a covariate because we suspect that many researchers make the decision to include a covariate only when and if the simpler analysis without such a covariate, the one they conduct first, is nonsignificant. So while there is nothing wrong with analyzing data with (or without) covariates, there is something wrong with analyzing it both ways and only reporting the one that works (Simmons, et al., 2011).

Some readers may not share our suspicions. Some readers may believe the inclusion of covariates is typically decided in advance, and that such studies do contain evidential value. Ultimately, this is an empirical question. If we are wrong, then *p*-curve should be right-skewed for experiments reporting only a covariate. If we are correct, then those studies should produce a *p*-curve that is uniform or, in the case of intense *p*-hacking, left-skewed.

Consistent with our hypothesis, Figure 3a shows that the *p*-curve for the set of studies reporting results only with a covariate lacks evidential value, rejecting the null of 33% power, $\chi^2(40)=82.5$, $p<.0001$. (It is also significantly *left*-skewed, $\chi^2(40) = 58.2$, $p=.031$).[14]

Figure 3b, in contrast, shows that *p*-curve for the other studies is significantly right-skewed, $\chi^2(42) = 94.2$, $p<.0001$, indicating that these studies do contain evidential value.[15] Nevertheless, it is

[14] A reviewer provided an alternative explanation for a left-skewed *p*-curve for this set of studies: some researchers who obtain both a low ANCOVA and ANOVA *p*-value for their studies choose to report the latter and hence exit our sample. We address this concern via simulations in Supplement 7, showing that this type of selection cannot plausibly account for the results presented in Figure 3A.

worth noting that (1) the observed *p*-curve closely resembles that of 33% power, (2) there is an uptick in *p*-curve at .05, and (3) three of the critical *p*-values in this set were greater than .05 (and hence excluded from *p*-curve).

*** Figure 3 ***

Selecting Studies

Analyzing a set of findings with *p*-curve requires selecting (1) the set of studies to include, and (2) the subset of *p*-values to extract from them. We discuss *p*-value selection in the next section. Here we propose four principles for study selection:

1) *Create a rule.* Rather than decide on a case-by-case basis whether a study should be included, one should minimize subjectivity by deciding on an inclusion rule in advance. The rule ought to be specific enough that an independent set of researchers could apply it and expect to arrive at the same set of studies.
2) *Disclose the selection rule*. Papers reporting *p*-curve analyses should disclose what the study selection rule was and justify it.
3) *Robustness to resolutions of ambiguity.* When the implementation of the rule generated ambiguity as to whether a given study should be included or not, results with and without those studies should be reported. This will inform readers of the extent to which the conclusions hinge on those ambiguous cases. For example if a *p*-curve is being constructed to examine the evidential value of studies of some manipulation on behavior, and it is unclear whether the dependent variable in a particular study does or does not qualify as behavior, the paper should report the results from *p*-curve with and without that study.

[15] There are 22 *p*-values included in Figure 3b because two studies involved reversing interactions, contributing two *p*-values each to *p*-curve. Both *p*-curves in Figure 3, then, are based on statistically significant *p*-values coming from 20 studies, our predetermined "sample size."

4) *Single-paper p-curve? Replicate.* One possible use of *p*-curve is to assess the evidential value of a single paper. This type of analysis does not lend itself to a meaningful inclusion rule. Given the risk of cherry-picking analyses that are based on single papers – for example, a researcher may decide to *p*-curve a paper precisely because s/he has already observed that it has many significant *p*-values greater than .025 – we recommend that such analyses be accompanied by a properly powered direct replication of at least one of the studies in the paper. Direct replications would enhance the credibility of results from *p*-curve, and impose a marginal cost that will reduce indiscriminate single-paper *p*-curving.

Peer-reviews of papers reporting *p*-curve analyses should ensure that the principles we have outlined above are followed or deviations from it be properly justified.

Selecting *p*-values

Most studies report multiple *p*-values, but not all *p*-values should be included in *p*-curve. Included *p*-values must meet three criteria: (1) test the hypothesis of interest, (2) have a uniform distribution under the null, and (3) be statistically independent of other *p*-values in *p*-curve.

Here we propose a five-step process for selecting *p*-values that meet these criteria (see Table 1). We refer to the authors of the original article as “researchers” and to those constructing the *p*-curve as “*p*-curvers.” It is essential for *p*-curvers to report how they implemented these steps. This can be easily achieved by completing our proposed standardized “*P*-curve Disclosure Table.” Table 2 consists of one such table; it reports a subset of the studies behind the right-skewed *p*-curve from Figure 3b. The *P*-curve Disclosure Table makes *p*-curvers accountable for decisions involved in creating a reported *p*-curve and facilitates discussion of such decisions. We strongly urge journals publishing *p*-curve analyses to require the inclusion of a *P*-curve Disclosure Table.

Step 1. Identify researchers' stated hypothesis and study design (Columns 1 and 2). As we discuss in detail in the next section, the researcher's stated hypothesis determines which *p*-values can and cannot be included in *p*-curve. *P*-curvers should report this first step by quoting, from the original paper, the stated hypothesis (Column 1). *P*-curvers should then characterize the study's design (Column 2).

Step 2. Identify the statistical result testing the stated hypothesis (Column 3). In Table 3 we identify the statistical result associated with testing the hypotheses of interest for the most common experimental designs (the next section provides a full explication of Table 3). Step 2 involves using Table 3 to identify and note the statistical result of interest.

Step 3. Report the statistical result(s) of interest (Column 4). P-curvers should quote directly from the relevant paragraph(s), or tables, in which results are reported.

Step 4. Recompute precise p-values based on reported test statistics (Column 5). We recommend recomputing precise *p*-values from reported test statistics. For example, if a paper reports "$F(1,76)=4.12$, $p<.05$" the *p*-curver should look up the *p*-value associated with $F=4.12$ for the $F(1,76)$ distribution. Recomputation is necessary because *p*-values are often reported merely as smaller than a particular benchmark (e.g., $p<.01$) and because they are sometimes reported incorrectly (Bakker & Wicherts, 2011). The online app available at http://p-curve.com does this automatically. The recomputed *p*-values should be reported in Column 5 of the disclosure table.

In some cases, authors do not report the results for the relevant statistical test. A common case is when the stated hypothesis involves the moderation of an effect but the authors report only simple effects and not the test associated with the interaction (Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011). If the results of interest cannot be computed from other information provided, then the *p*-value of interest cannot be included in *p*-curve. The *p*-curver should report this in Column 5 (e.g., by writing “Not Reported”).

Step 5. Report robustness p-values (Column 6). Some experiments report results on two or more correlated dependent variables (e.g., how much people like a product and how much they are willing to pay for it). This will lead to multiple *p*-values associated with a given finding. Similarly, sometimes researchers will report multiple analyses on the same measure (e.g., differences in liking of the product across conditions tested with both a *t*-test and a non-parametric test). *P*-curvers should not simultaneously include all of these *p*-values because they must be statistically independent for inference from *p*-curve to be valid. Instead, *p*-curvers should use selection rules and report robustness of the results to such rules. For example, report a *p*-curve for the first *p*-value that is reported across all studies, and another for the last one that is reported.

P-curves should document this last step by quoting, from the original paper, all statistical analyses that produced such *p*-values (in Column 4), and report the additional *p*-values (in Column 6). See last row of our Table 2.

*** Tables 1& 2 ***

P-curving findings that do not test the researcher’s stated hypothesis. We envision that most applications of *p*-curve will involve assessing the evidential value of the core findings put forward in a set of studies, whatever those findings were. Examples include assessing the evidential value of findings aggregated by paper, author, journal, or method of analysis. For instance, if a *p*-curver were interested in assessing the evidential value of the Top 10 most cited empirical articles published in this journal, her statistical results of interest would be those testing the stated hypotheses in those articles, whatever those hypotheses might be.

Sometimes, however, *p*-curve will be used to aggregate studies based on a *finding*, as in traditional meta-analytic applications, and that finding of interest to the *p*-curver may not be the one directly testing the stated hypothesis by the original researchers. Thus, it may not correspond to that arising from looking up the study design in Table 3.

For example, a *p*-curver may be interested in meta-analyzing the evidential value of gender differences in helping behavior, but may be obtaining *p*-values from papers whose authors' stated hypothesis pertains to the effects of mood on helping behavior and who just happened to also report the impact of gender. Similarly, a *p*-curver may be interested in a simple two-cell comparison comprising a subset of a 2x2x2 design. When this occurs, a column should be added to the *P*-curve Disclosure Table, Column 3.5, placed between Columns 3 and 4, identifying the finding to be *p*-curved.

In some of those cases, when the finding of interest to the *p*-curver is not that identified by Table 3, the result may not validly be included in *p*-curve even if it is reported in the original paper. For example, if a *p*-curver is interested in the relationship between two variables (e.g., gender and math performance) but the researcher's interest was to investigate a moderator of this relationship (e.g., showing that this relationship is weaker under some condition), then the finding of interest to the *p*-curver (the simple effect) cannot be included in *p*-curve because when researchers investigate attenuated interactions, *p*-values for simple effects are not uniformly distributed under the null.

This is explained more fully in the next section; the important lesson at this juncture is that when *p*-curvers are interested in findings different from those testing the stated hypothesis, and thus different from those indicated in Table 3, they may not be able to validly include the study in *p*-curve.

Selecting *p*-values for specific study designs

For simple designs, such as two- or three-cell designs, identifying which *p*-value is valid to select is simple (see Table 3). However, many psychological findings involve interactions. Interactions occur when the impact of predictor X on dependent variable Y is influenced by a third variable Z. For example, the impact of winter vs. summer (X) on sweating (Y) can be influenced by whether the person is indoors vs. outdoors (Z).

When the researcher's stated hypothesis is that the interaction *attenuates* the impact of X on Y (e.g., people always sweat more in summer, but less so indoors), the relevant test is whether the interaction is significant (Gelman & Stern, 2006), and hence *p*-curve must include *only* the interaction's *p*-value. Even if the *p*-curver is only interested in the simple effect (e.g., because she is conducting a meta-analysis on this effect), that *p*-value may not be included in *p*-curve if it was reported as part of an attenuated interaction design. *Simple effects from a study examining the attenuation of an effect should not be included in p-curve, as they bias p-curve to conclude evidential value is present even when it is not.*[16]

When the researcher's stated hypothesis is that the interaction *reverses* the impact of X on Y (e.g., people sweat more outdoors in the summer, but more indoors in the winter), the relevant test is whether the two simple effects of X on Y are of opposite sign *and* are significant, and so both simple effects' *p*-values ought to go into *p*-curve. The interaction that is predicted to reverse the

[16] When a researcher investigates the attenuation of an effect, the interaction term will tend to have a larger *p*-value than the simple effect (because the latter is tested against a null of 0 with no noise, and the former against another parameter estimated with noise). Because publication incents the *interaction* to be significant, *p*-values for the simple effect will usually have to be much smaller than .05 in order for the study to be published. This means that, for the simple effect, even *p*-values below .05 will be censored from the published record and, thus, not uniformly distributed under the null.

sign of an effect should not be included in *p*-curve, as it biases *p*-curve to conclude evidential value is present even when it is not.[17]

In some situations it may be of interest to assess the evidential value of only one of the opposing simple effects for a predicted reversal (e.g., if one is obvious and the other counterintuitive). As long as a general rule is set for such selections and shared with readers, subsets of simple effect *p*-values may validly be selected into *p*-curve as both are distributed uniform under the null.

Sometimes one may worry that the authors of a study modify their "predictions" upon seeing the data (Kerr, 1998), such that an interaction is reported as predicted to reverse sign only because this is what occurred. Changing a hypothesis to fit the data is a form of *p*-hacking. Thus, the rules outlined above still apply in those circumstances, as we can rely on *p*-curve to assess if this or any other form of *p*-hacking have undermined the evidential value of the data. Similarly, if a study predicts an attenuated effect, but the interaction happens to reverse it, the *p*-value associated with testing the hypothesis of interest is still the interaction. Keep in mind that a fully attenuated effect (d=0) will be estimated as negative half the time.

For more complicated designs, for example a 2x3, one can easily extend the logic above. For example, consider a 2x3 design where the relationship between an ordinal, 3-level (i.e., with levels *low*, *medium*, and *high*) is predicted to reverse with the interaction (e.g., people who practice more for a task in the summer perform better, but people who practice more in the winter perform worse; see Table 3). Because here an effect is predicted to *reverse*, *p*-curve ought to include both "simple" effects, which in this case would be both linear trends. If the prediction were that in winter

[17] Similar logic to the previous footnote applies to studies examining reversals. One can usually not publish a reversing interaction without getting both simple effects to be significant (and, by definition, opposite). This means that the reversing interaction *p*-value will necessarily be much smaller than .05 (and thus some significant reversing interactions will be censored from the published record), and thus not uniformly distributed under the null.

the effect is merely attenuated rather than reversed, *p*-curve would then include the *p*-value of the attenuated effect: the interaction of both trends.

Similarly, if a 2x2x2 design tests the prediction that a two-way interaction is itself attenuated, then the *p*-value for the three-way interaction ought to be selected. If it tests the prediction that a two-way interaction is reversed, then the *p*-values of both two-way interactions ought to be selected.

Table 3

How often does *p*-curve get it wrong?

We have proposed using *p*-curve as a tool for assessing if a set of statistically significant findings contains evidential value. Now we consider how often *p*-curve gets these judgments right and wrong. The answers depend on (1) the number of studies being *p*-curved, (2) their statistical power, and (3) the intensity of *p*-hacking. Figure 4 reports results for various combinations of these factors for studies with a sample size of *n*=20 per cell.

*** Figure 4***

First, let us consider *p*-curve's power to detect evidential value: How often does *p*-curve correctly conclude that a set of real findings in fact contains evidential value (i.e., that the test for right skew is significant)? Figure 4A shows that with just five *p*-values, *p*-curve has more power than the individual studies on which it is based. With 20 *p*-values, it is virtually guaranteed to detect evidential value, even when the set of studies is powered at just 50%.

Figure 4B considers the opposite question: How often does *p*-curve incorrectly conclude that a set of real findings lack evidential value (i.e., that *p*-curve is significantly less right-skewed than when studies are powered at 33%)? With a significance threshold of .05, then by definition this probability is 5% when the original studies were powered at 33% (because when the null is true, there is a 5% chance of obtaining *p*<.05). If the studies are more properly powered, this rate drops accordingly. Figure 4B shows that with as few as 10 *p*-values, *p*-curve almost never falsely concludes that a set of properly powered studies – a set of studies powered at about 80% - lacks evidential value.

Figure 4C considers *p*-curve's power to detect lack of evidential value: How often does *p*-curve correctly conclude that a set of false-positive findings lack evidential value (i.e., that the test for 33% power is significant)? The first set of bars show that, in the absence of *p*-hacking, 62% of *p*-curves based on 20 false-positive findings will conclude the data lack evidential value. This probability rises sharply as the intensity of *p*-hacking increases.

Finally, Figure 4D considers the opposite question: How often does *p*-curve falsely conclude that a set of false-positive findings contain evidential value (i.e., that the test for right skew is significant)? By definition, the false-positive rate is 5% when the original studies do not contain any *p*-hacking. This probability drops as the frequency or intensity of *p*-hacking increases. It is virtually impossible for *p*-curve to erroneously conclude that 20 intensely *p*-hacked studies of a nonexisting effect contain evidential value.

In sum, when analyzing even moderately powered studies, *p*-curve is highly powered to detect evidential value; and when a nonexistent effect is even moderately *p*-hacked, it is highly powered to detect lack of evidential value. The "false-positive" rates, in turn, are often *lower* than nominally reported by the statistical tests because of the conservative nature of the assumptions underlying them.

Cherry-picking *p*-curves

All tools can be used for harm and *p*-curve is no exception. For example, a researcher could cherry-pick high *p*-value studies by area, journal, or set of researchers, and then use the resulting *p*-curve to unjustifiably argue that those data lack evidential value. Similarly, a researcher could run *p*-curve on paper after paper, or literature after literature, and then report the results of the significantly left-skewed *p*-curves without properly disclosing how the set of *p*-values was arrived at. This is not just a hypothetical concern, as tests for publication bias have been misused in just this manner (for the discussion of one such case see Simonsohn, 2012, 2013). How concerned should we be about the misuse of *p*-curve? How can we prevent it? We address these questions below.

How concerned should we be with cherry-picking? Not too much. As was shown in Figure 4, *p*-curve is unlikely to lead one to falsely conclude that data lack evidential value, even when a set of studies is just moderately powered. For example, a *p*-curve of 20 studies powered at just 50% has less than a 1% chance of concluding the data lack evidential value (rejecting the 33% power test). This low false-positive rate already suggests it would be difficult to be a successful cherry-picker of *p*-curves.

We performed some simulations to consider the impact of cherry picking more explicitly. We considered a stylized world in which papers have four studies (i.e., *p*-values), the cherry-picker targets 10 papers, and chooses to report *p*-curve on the worst five of the ten papers. The cherry-picker, in other words, ignores the most convincing five papers, and writes her critique on the least convincing five. What will the cherry-picker find?

The results of 100,000 simulations of this situation showed that if studies in the targeted set were powered at 50%, the cherry-picker would almost *never* conclude that *p*-curve is left-skewed, only 11% of the time conclude that it lacks evidential value (rejecting 33% power), and 54% of the time correctly conclude that it contains evidential value. The remaining 35% of *p*-curves would be inconclusive. If the studies in that literature were properly powered at 80%, then just about all (>99.7%) of the cherry-picked *p*-curves are expected to correctly conclude the data have evidential value.

How to prevent cherry-picking? Disclosure. In the Selecting Studies and *P*-values sections, we provided guidelines for disclosing such selection. As we have argued elsewhere, disclosure of ambiguity in data collection and analysis can dramatically reduce the impact of such ambiguity on false-positives (Simmons, et al., 2011; Simmons, Nelson, & Simonsohn, 2012). This applies as much to disclosing how sample size was determined for a particular experiment, and what happens if one does not control for a covariate, as it does for describing how one determined which papers belonged in a literature and what happens if certain arbitrary decisions are reversed. If journals publishing *p*-curve analyses require authors to report their rules for study selection and their *P*-curve Disclosure Tables, it becomes much more difficult to misuse *p*-curve.

***P*-curve vs. Other Methods**

In examining *p*-value distributions for a set of statistically significant findings, *p*-curve is related to previous work that has examined the distribution of *p*-values (or their corresponding t or Z scores) reported across large numbers of articles (Card & Krueger, 1995; Gadbury & Allison, 2012; Gerber & Malhotra, 2008a, 2008b; Masicampo & Lalande, 2012; Ridley et al., 2007). Because these papers (1) include *p*-values arising from all reported tests (not just those associated with the finding of interest), and/or (2) focus on discrete differences around cutoffs points (instead of differences between overall observed and expected distributions under a null), they do not share *p*-curve's key contribution: the ability to assess if a set of statistically significant findings contains evidential value.

Three main approaches exist for addressing selective reporting. The most common uses "funnel plots," concluding that publication bias is present if reported effect sizes correlate with sample sizes (Duval & Tweedie, 2000; Egger et al., 1997). A second approach is the "fail safe" method. It assesses the number of unpublished studies that would need to exist to make an overall effect size statistically (Rosenthal, 1979) or practically (Orwin, 1983) non-significant. A third and more recent approach is the “excessive-significance test.” It assesses whether the share of significant findings is higher than that implied by the statistical power of the underlying studies (Ioannidis & Trikalinos, 2007).

All three of these approaches suffer from important limitations that *p*-curve does not suffer from. First, when true effect sizes differ across studies, as they inevitably do, the funnel plot and the excessive significance approaches risk falsely concluding publication bias is present when in fact it is not (Lau et al., 2006; Peters et al., 2007; Tang & Liu, 2000). Second, whereas *p*-curve assesses the *consequences* of publication bias, neither the fail-safe method nor the excessive-significance test do so. The funnel plot’s fill-and-trim procedure does, but it is only valid if publication bias is affected only by effect size, not by statistical significance (Duval & Tweedie, 2000), even though in many disciplines, including psychology, publication bias operates primarily through statistical significance rather than effect size (Fanelli, 2012; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). Third, as we mentioned in the introduction, *p*-hacking invalidates the reassurances provided by the fail-safe method. With enough *p*-hacking, a large set of false-positives can be achieved with file-drawers free of failed studies. In short, only *p*-curve provides an answer to the question, “Does this set of statistically significant findings contain evidential value?”

Limitations

Two main limitations of *p*-curve , in its current form, are that it does not yet technically apply to studies analyzed using discrete test statistics (e.g., difference of proportions tests) and that is less likely to conclude data have evidential value when a covariate correlates with the variable of interest (e.g., in the absence of random assignment).

Formally extending *p*-curve to discrete tests (e.g., contingency tables) imposes a number of technical complications that we are attempting to address in ongoing work, but simulations suggest that the performance of *p*-curve on difference of proportions tests treated as if they were continuous (i.e., ignoring the problem) leads to tolerable results (see Supplement 4).

Extending *p*-curve to studies with covariates correlated with the independent variable requires accounting for the effect of collinearity on standard errors and hence *p*-values. One could use *p*-curve as a conservative measure of evidential value in non-experimental settings (*p*-curve is biased flat).

Another limitation is that *p*-curve will too often fail to detect studies that lack evidential value. First, because *p*-curve is markedly right-skewed when an effect is real but only mildly left-skewed when a finding is *p*-hacked, if a set of findings combine (a few) true effects with (several) nonexistent ones, *p*-curve will usually not detect the latter. Something similar occurs if studies are confounded. For example, findings showing that egg consumption has health benefits may produce a markedly right-skewed *p*-curve if those assigned to consume eggs were also asked to exercise more. In addition, *p*-curve does not include any *p*-values above .05, even those quite close to significance (e.g., *p*=.051). This excludes *p*-values that would be extremely infrequent in the presence of a true effect, and therefore diagnostic of a nonexistent effect.

Why focus on *p*-values?

Significance testing has a long list of weaknesses and a longer list of opponents. Why propose a new methodology based on an old-fashioned and often-ridiculed statistic? We care about *p*-values because researchers care about *p*-values. Our goal is to partial out the distorting effect of researcher behavior on scientific evidence seeking to obtain .05; because that behavior is motivated by *p*-values, we can use *p*-values to undo it. If the threshold for publication was based on posteriordistributions, some researchers would engage in posterior-hacking, and one would need to devise a posterior-curve to correct for it.

It is worth noting that while Bayesian statistics are not generally influenced by data peeking (ending data collection upon obtaining a desired result), they are influenced by other forms of *p*-hacking. In particular, Bayesian inference also exaggerates the evidential value of results obtained with undisclosed exclusions of measures, participants or conditions, or by cherry-picked model assumptions and specifications. Switching from frequentist to Bayesian statistics has many potential benefits, but eliminating the impact of the self-serving resolution of ambiguity in data collection and analysis is not one of them.

Conclusions

Selective reporting of studies and analyses is, in our view, an inevitable reality, one that challenges the value of the scientific enterprise. In this paper we have proposed a simple technique to bypass some of its more serious consequences on hypothesis testing. We show that examining the distribution of significant *p*-values, *p*-curve, one can assess whether selective reporting can be rejected as the sole explanation for a set of significant findings.

P-curve has high power to detect evidential value even when individual studies are underpowered. Erroneous inference from *p*-curve is unlikely, even when it is misused to analyze a cherry-picked set of studies. It can be applied to diverse set of findings to answer questions for which no existing technique is available. While *p*-curve may not lead to a reduction of selective reporting, at present it does seem to provide the most flexible, powerful, and useful tool for taking into account the impact of selective reporting on hypothesis testing.

References

Bakker, M., & Wicherts, J. M. (2011). The (Mis) Reporting of Statistical Results in Psychology Journals. *Behavior research methods, 43*(3), 666-678.

Bross, I. D. J. (1971). Critical Levels, Statistical Language and Scientific Inference. *Foundations of Statistical Inference*, 500-513.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally.

Card, D., & Krueger, A. B. (1995). Time-Series Minimum-Wage Studies: A Meta-Analysis. *The American Economic Review*, *85*(2), 238-243.

Clarkson, J. J., Hirt, E. R., Jia, L., & Alexander, M. B. (2010). When Perception Is More Than Reality: The Effects of Perceived Versus Actual Resource Depletion on Self-Regulatory Behavior. *Journal of Personality and Social Psychology*, *98*(1), 29.

Cohen, J. (1962). The Statistical Power of Abnormal-Social Psychological Research: A Review. *Journal of Abnormal and Social Psychology, 65*(3), 145-153.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*: Lawrence Erlbaum.

Cohen, J. (1992). A Power Primer. *Psychological bulletin, 112*(1), 155.

Cohen, J. (1994). The Earth Is Round (P<. 05). *American Psychologist, 49*(12), 997.

Cole, L. C. (1957). Biological Clock in the Unicorn. *Science, 125*(3253), 874-876.

Cumming, G. (2008). Replication and P Intervals: P Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science, 3*(4), 286-300.

Duval, S., & Tweedie, R. (2000). Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics*, *56*(2), 455-463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in Meta-Analysis Detected by a Simple, Graphical Test. *Bmj, 315*(7109), 629-634.

Fanelli, D. (2012). Negative Results Are Disappearing from Most Disciplines and Countries. *Scientometrics, 90*(3), 891-904.

Gadbury, G. L., & Allison, D. B. (2012). Inappropriate Fiddling with Statistical Analyses to Obtain a Desirable P-Value: Tests to Detect Its Presence in Published Literature. *PLoS ONE, 7*(10), e46363. doi: 10.1371/journal.pone.0046363

Gelman, A., & Stern, H. (2006). The Difference between "Significant" and "Not Significant" Is Not Itself Statistically Significant. *The American Statistician, 60*(4), 328-331.

Gerber, A. S., & Malhotra, N. (2008a). Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science, 3*(3), 313-326.

Gerber, A. S., & Malhotra, N. (2008b). Publication Bias in Empirical Sociological Research Do Arbitrary Significance Levels Distort Published Results? *Sociological Methods & Research, 37*(1), 3-30.

Greenwald, A. G. (1975). Consequences of Prejudice against the Null Hypothesis. *Psychological Bulletin; Psychological Bulletin, 82*(1), 1.

Hodges, J., & Lehmann, E. (1954). Testing the Approximate Validity of Statistical Hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological),* 261-268.

Hung, H. M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The Behavior of the P-Value When the Alternative Hypothesis Is True. *Biometrics,* 11-22.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. [Editorial Material]. *Plos Medicine, 2*(8), 696-701.

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology, 19*(5), 640-646.

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An Exploratory Test for an Excess of Significant Findings. *Clinical Trials, 4*(3), 245-253.

Kerr, N. L. (1998). Harking: Hypothesizing after the Results Are Known. *Personality and Social Psychology Review, 2*(3), 196-217.

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin, 108*(3), 480-498.

Lau, J., Ioannidis, J., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The Case of the Misleading Funnel Plot. *BMJ, 333*(7568), 597.

Lehman, E. L. (1986). *Testing Statistical Hypotheses* (2nd ed.): Wiley

Masicampo, E., & Lalande, D. R. (2012). A Peculiar Prevalence of P Values Just Below. 05. *The Quarterly Journal of Experimental Psychology, 65*(11), 2271-2279.

Nelson, L. D., Simonsohn, U., & Simmons, J. P. (2013). *P-Curve Estimates Publication Bias Free Effect Sizes*.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous Analyses of Interactions in Neuroscience: A Problem of Significance. *Nature Neuroscience, 14*(9), 1105-1107.

Orwin, R. G. (1983). A Fail-Safe N for Effect Size in Meta-Analysis. *Journal of Educational Statistics*, 157-159.

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science, 7*(6), 531-536.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the Trim and Fill Method in the Presence of Publication Bias and between-Study Heterogeneity. *Statistics in medicine, 26*(25), 4544-4562.

Phillips, C. V. (2004). Publication Bias in Situ. *BMC Medical Research Methodology, 4*(1), 20.

Ridley, J., Kolm, N., Freckelton, R., & Gage, M. (2007). An Unexpected Influence of Widely Used Significance Thresholds on the Distribution of Reported P-Values. *Journal of evolutionary biology, 20*(3), 1082-1089.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin, 86*(3), 638.

Serlin, R., & Lapsley, D. (1985). Rationality in Psychological Research: The Good-Enough Principle. *The American psychologist, 40*(1), 73-83.

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*(11), 1359-1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 Word Solution. *Dialogue, The Official Newsletter of the Society for Personality and Social Psychology, 26*(2), 4-7.

Simonsohn, U. (2012). It Does Not Follow Evaluating the One-Off Publication Bias Critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in Press). *Perspectives on Psychological Science, 7*(6), 597-599.

Simonsohn, U. (2013). It Really Just Does Not Follow, Coments on Francis (2013). *Journal of Mathematical Psychology*.

Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance--or Vice Versa. *Journal of the American statistical association*, 30-34.

Sterling, T. D., Rosenbaum, W., & Weinkam, J. (1995). Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *American Statistician*, 108-112.

Tang, J. L., & Liu, J. L. Y. (2000). Misleading Funnel Plot for Detection of Bias in Meta-Analysis. *Journal of Clinical Epidemiology, 53*(5), 477-484.

Topolinski, S., & Strack, F. (2010). False Fame Prevented: Avoiding Fluency Effects without Judgmental Correction. *Journal of Personality and Social Psychology, 98*(5), 721.

Van Boven, L., Kane, J., McGraw, A. P., & Dale, J. (2010). Feeling Close: Emotional Intensity Reduces Perceived Psychological Distance. *Journal of Personality and Social Psychology, 98*(6), 872.

Wallis, W. A. (1942). Compounding Probabilities from Independent Significance Tests. *Econometrica*, 229-248.

White, H. (2003). A Reality Check for Data Snooping. *Econometrica, 68*(5), 1097-1126.

Wohl, M. J., & Branscombe, N. R. (2005). Forgiveness and Collective Guilt Assignment to Historical Perpetrator Groups Depend on Level of Social Category Inclusiveness. *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology, 88*(2), 288.

Table 1. Steps for selecting *p*-values and creating *P*-curve Disclosure Table

Step 1	Identify researchers' stated hypothesis and study design	(Columns 1 &2)
Step 2	Identify the statistical result testing stated hypothesis (use Table 3)	(Column 3)
Step 3	Report the statistical results of interest	(Column 4)
Step 4	Recompute the precise *p*-value(s) based on reported test statistics	(Column 5)
Step 5	Report robustness results	(Column 6)

Table 2. Sample *P*-curve Disclosure Table

Original Paper *(Study 1 of each paper)*	**(1) Quoted text from original paper indicating prediction of interest to researchers**	**(2) Study design**	**(3) Key statistical result** *(looking up column 2 in table 3)*	**(4) Quoted text from original paper with statistical results**	**(5) Results**	**(6) Robustness results**
Van Boven et al. (2010)	We predicted that people would perceive their embarrassing moment as less psychologically distant when described emotionally.	**two-cell (description: emotional vs. not)**	**Difference of means**	As predicted, participants perceived their previous embarrassing moment to be less psychologically distant after describing it emotionally (*M* = 4.90, *SD* = 2.30) than after describing it neutrally (*M* = 6.66, *SD* = 1.83), ***t* (38) = 2.67, *p* < .025** (see Table 1).	t(38)=2.67, p=0.0111	
Topolinksi & Strack (2010)	We predicted that the classical effect by Jacoby, Kelley, et al., namely the **misattribution** of increased fluency to fame, **would vanish** under the oral motor task but would still be detected under a manual motor task.	**2 (exposure: old vs. new) x** **2 (motor task: oral vs manual)** *(attenuated interaction)*	**Two-way interaction**	Over the fame ratings in the test phase, a 2 (exposure: old items, new items) × 2 (concurrent motor task: manual, oral) analysis of variance (ANOVA) was run with motor task as a between-subjects factor. A main effect of exposure, *F* (1, 48) = 5.54, *p* < .023, ηp2 = .10, surfaced, as well as an interaction between exposure and motor task, ***F* (1, 48) = 4.12, *p* < .05, ηp2 = .08.** The conditional means are displayed in Table 1.	F(1,48)=4.12, p=0.0479	
Clarkson et al. (2010)	Specifically, participants in the low depletion condition were expected [...] to **persist longer** on our problem-solving task when given the replenished (vs. depleted) feedback. **Conversely,** participants in the high depletion condition were expected [...] to persist longer on our problem-solving task when given the depleted (vs. replenished) feedback.	**2 (depletion: high vs low) x** **2 (feedback: depleted vs. replenished)** *(reversing interaction)*	**Two simple effects**	In the low depletion condition, participants persisted significantly longer when given the replenished, as opposed to depleted, feedback, ***t* (30) = –2.52, *p* < .02**. In the high depletion condition, participants persisted significantly longer when given the depleted, as opposed to replenished, feedback, ***t* (30) = 2.50, *p* < .02.**	t(30)=2.52, p=0.0173 t(30)=2.5, p=0.0181	
Wohl & Branscombe (2005)	We expected that Jews would be more willing to **forgive** Germans for the past when they categorized at the human identity level and that the **guilt assigned** to contemporary Germans would be lower in the human identity condition compared with the social identity condition.	**two-cell (identity: human vs. social)**	**Difference of means** *(for two d.v.s)*	Participants assigned significantly less collective guilt to Germans when the more inclusive human-level categorization was salient (M = 5.47, SD = 2.06) than they did when categorization was at the social identity level (M = 6.75, SD = 0.74), **F(1, 45) = 7.62, p <.01**, d = 0.83. Participants were more willing to **forgive** Germans when the human level of identity was salient (M = 5.84, SD = 1.25) than they were when categorization was at the social identity level (M = 4.52, SD = 0.92)**, F(1, 45) = 16.55, p <.01, d = 1.20.**	F(1,45)=7.62, p=0.0083	F(1,45)=16.55, p=0.0002

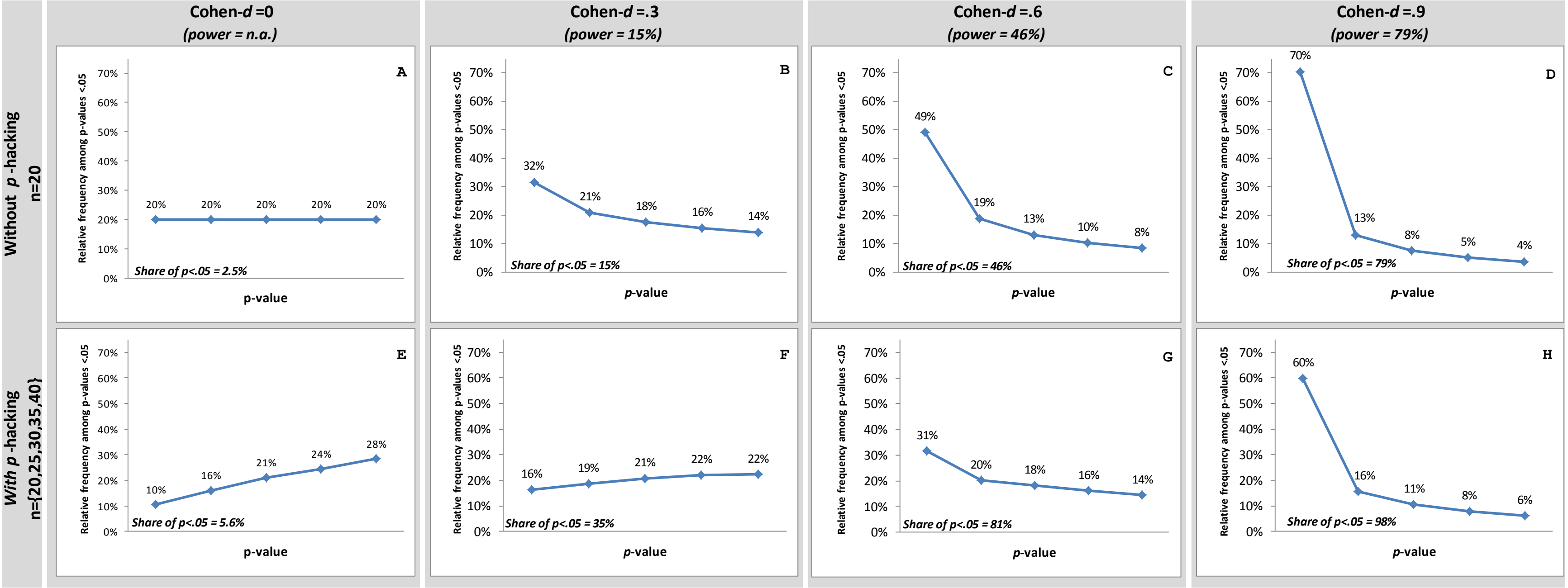
Notes: This table includes a subset of four of the twenty studies used in our demonstration from Figure 3b. The full table is reported in Supplement 5.

Table 3. Which *p*-values to select from common study designs.

DESIGN	EXAMPLE	WHICH RESULT TO INCLUDE	
		IN MAIN P-CURVE	IN ROBUSTNESS TEST
3-Cell	*Examining how math training affects math performance*		
High **Medium** **Low**	60 minutes of math training 30 minutes of math training 5 minutes of math training	Linear trend	High vs. low comparison
Treatment **Control 1** **Control 2**	60 minutes of math training 60 minutes of unrelated training No training	Treatment vs. Control 1	Treatment vs. control 2
Treatment 1 **Treatment 2** **Control**	60 minutes of math training, start with easy questions 60 minutes of math training, start with hard questions No training	Treatment 1 vs. Control	Treatment 2 vs. Control
2X2 DESIGN	*Examining how season interacts with being indoors vs. outdoors to affect sweating*		
Attenuated Interacton	Always sweat more in summer, but less so when indoors.	2x2 Interaction	
Reversing Interacton	Sweat more in summer than winter when outdoors, but more in winter than in summer when indoors	Both simple effects	
3x2 DESIGN	*Examining how season interacts with math training to affect math performance*		
Attenuated Trends	More math training (60 vs. 30 vs. 5 minutes) leads to better performance always, but more so in winter than in summer	Difference in linear trends	2x2 interaction for high vs. low
Reversing Trends	More math training (60 vs. 30 vs. 5 minutes) leads to better performance in winter, but worse performance in summer	Both linear trends	Both high vs. low comparisons
2x2x2 DESIGN	*Examining how gender and season interact with being indoors vs. outdoors to affect sweating*		
***Attenuation* of attenuated interaction**	Both men and women sweat more in summer than winter, but less so when indoors. This attenuation is stronger for men than for women.	Three-way interaction	
***Reversal* of attenuated interaction**	Men sweat more in summer than winter, but less so when indoors. Women also sweat more in summer than winter, but more so when indoors.	Both two-way interactions	
***Reversal* of reversing interaction**	Men sweat more in summer than winter when outdoors, but more in winter than in summer when indoors. Women sweat more in winter than summer when outdoors, but more in summer than winter when indoors.	All four simple effects	

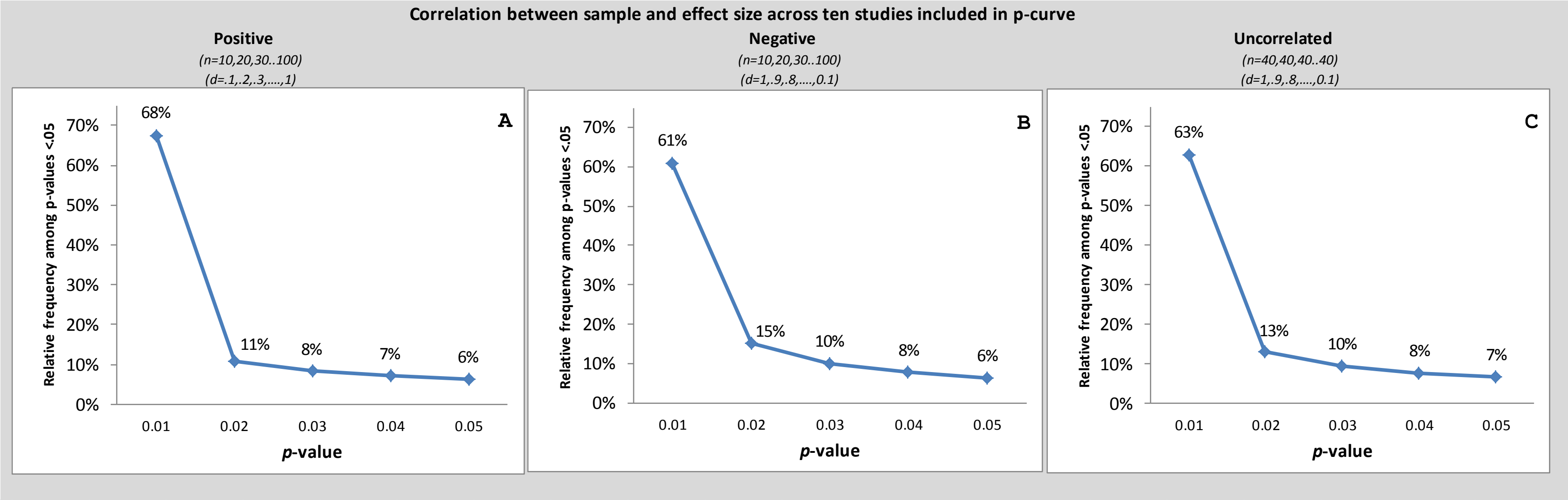
Figure Captions

Figure 1. P-curves for different true effect sizes in the presence and absence of p-hacking.



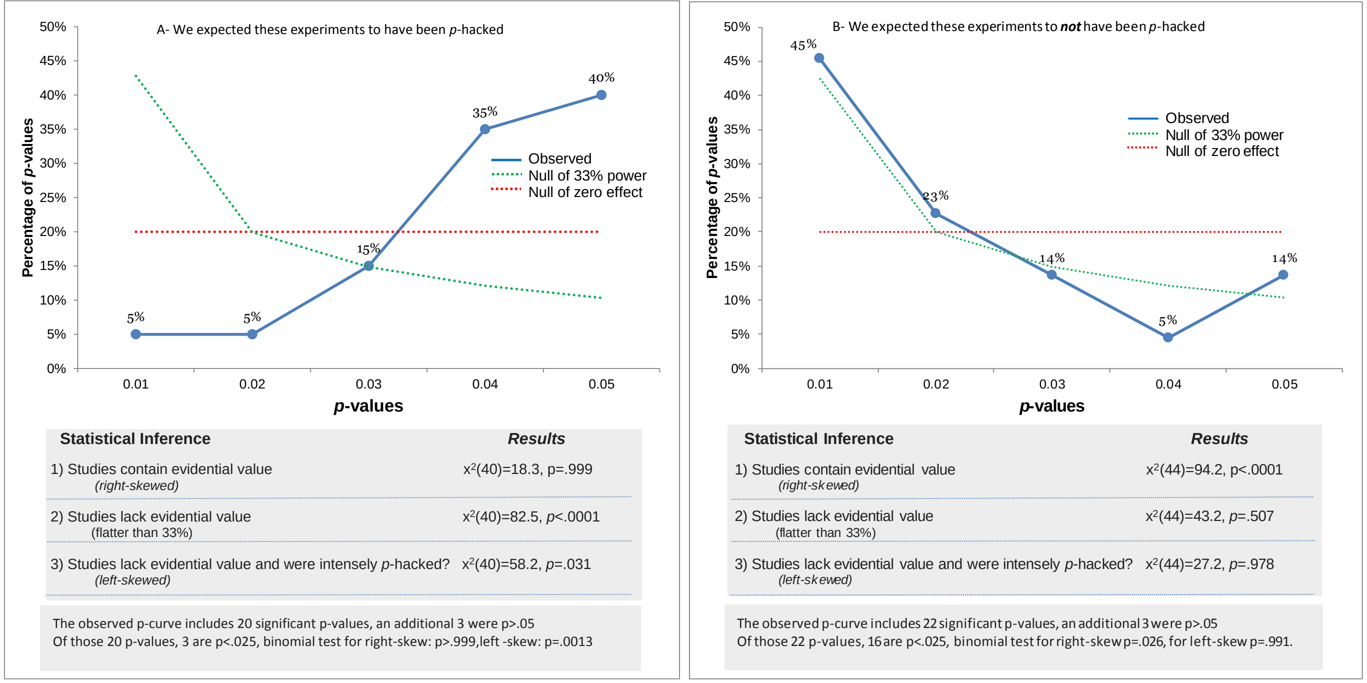
Note. Graphs depict expected *p*-curves for difference-of-means *t*-tests for samples from populations with means differing by *d* standard deviations. A through D are products of the central and noncentral *t*-distribution (see Supplement 1). E through H are products of 400,000 simulations of two samples with 20 normally distributed observations. For E through H, if the difference was not significant, 5 additional independent observations were added to each sample, up to a maximum of 40 total observations. “Share of *p*<.05” indicates the share of all studies producing a statistically significant effect using a two-tailed test for a directional prediction (hence 2.5% under the null).

***Figure 2*. Expected *p*-curves are right-skewed for sets of studies containing evidential value, no matter the correlation between the studies' effect sizes and sample sizes.**



Note. Graphs depict expected *p*-curves for difference-of-means t-tests for 10 studies for which effect size (d) and sample size (n) are perfectly positively correlated (A), perfectly negatively correlated (B), or uncorrelated (C). Per-condition sample sizes ranged from 10 to 100, in increments of 10, and effect sizes ranged from .10 to 1.00, in increments of .10. For example, for the positive correlation case (A), the smallest sample size (*n* = 10) had the smallest effect size (*d* = .10); for the negative correlation case (B), the smallest sample size (n = 10) had the largest effect size (d = 1.00); and, for the uncorrelated case (C), the effect sizes ranged from .10 to 1.00 while the sample sizes remained constant (n = 40). Expected *p*-curves for each of the sets of ten studies were obtained from noncentral distributions (Supplement 1) and then averaged across the ten studies. For example, to generate Panel A, we determined that 22% of significant *p*-values will be less than .01 when n = 10 and d = .10; 27% will be less than .01 when n = 20 and d = .20; 36% will be less than .01 when n = 30 and d = .30; and so on. Panel A shows that, when one averages these percentages across the ten possible combinations of effect sizes and sample sizes, 68% of significant *p*-values will be below .01. Similar calculations were performed to generate Panels B and C

***Figure 3*. *P*-curves for JPSP studies suspected to have been *p*-hacked (A) and not *p*-hacked (B).**



Statistical Inference	*Results*
1) Studies contain evidential value *(right-skewed)*	x²(40)=18.3, p=.999
2) Studies lack evidential value (flatter than 33%)	x²(40)=82.5, *p*<.0001
3) Studies lack evidential value and were intensely *p*-hacked? *(left-skewed)*	x²(40)=58.2, *p*=.031

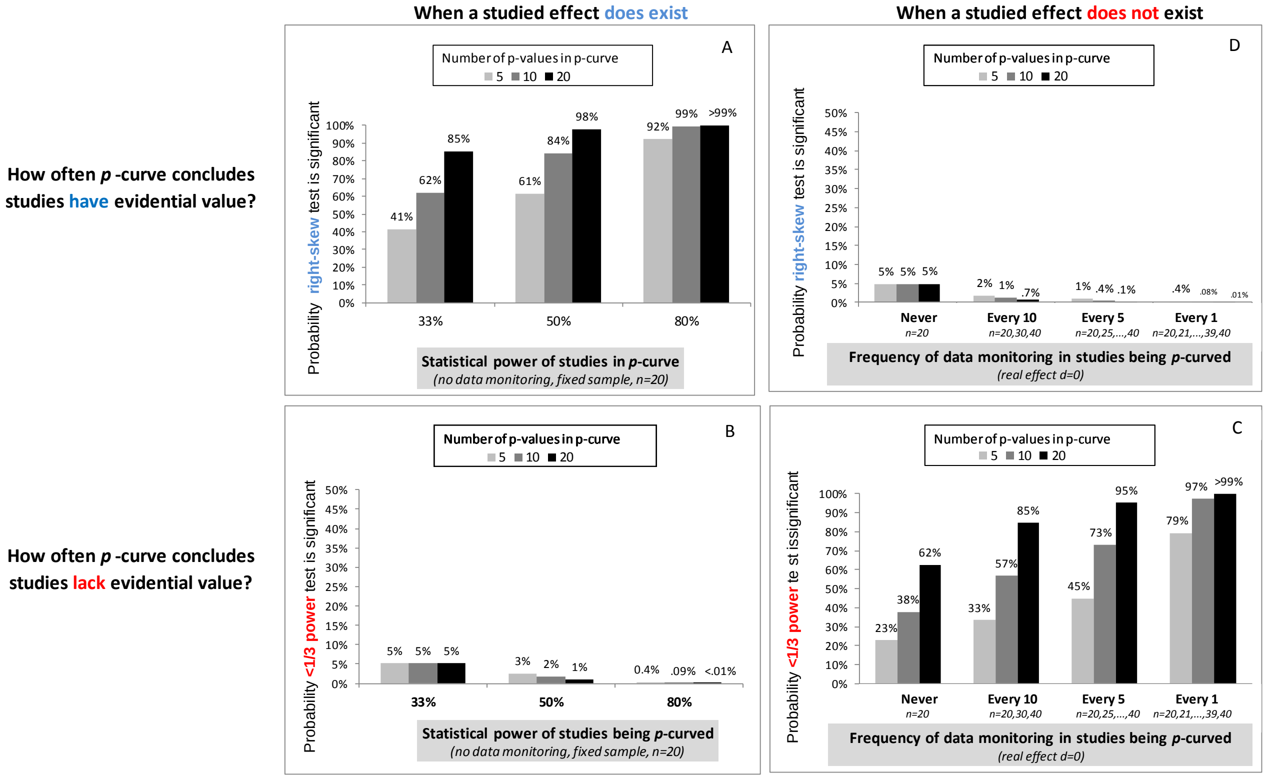
The observed p-curve includes 20 significant p-values, an additional 3 were p>.05
Of those 20 p-values, 3 are p<.025, binomial test for right-skew: p>.999,left -skew: p=.0013

Statistical Inference	*Results*
1) Studies contain evidential value *(right-skewed)*	x²(44)=94.2, p<.0001
2) Studies lack evidential value (flatter than 33%)	x²(44)=43.2, *p*=.507
3) Studies lack evidential value and were intensely *p*-hacked? *(left-skewed)*	x²(44)=27.2, *p*=.978

The observed p-curve includes 22 significant p-values, an additional 3 were p>.05
Of those 22 p-values, 16 are p<.025, binomial test for right-skew p=.026, for left-skew p=.991.

Notes: Graphs depict *p*-curves observed in two separate sets of 20 studies. The first set (A) consists of 20 JPSP studies that only report statistical results, from an experiment with random assignment, controlling for a covariate; we suspected this indicated *p*-hacking. The second set (B) consists of 20 JPSP studies reported in papers whose entire full text does not include keywords we suspected could indicate *p*-hacking (e.g., “exclude”, “covariate”).

Figure 4. How often *p*-curve correctly and incorrectly diagnoses evidential value.



Note. The bars indicate how often *p*-curve would lead one to conclude that a set of findings contains evidential value (a significant right-skew; Panels A & D) or does not contain evidential value (powered significantly below 33%; Panels B & C). Results are based on 100,000 simulated *p*-curves. For Panels A and B, the simulated *p*-curves are derived from *p*-values drawn at random from noncentral distributions. For Panels C and D, the *p*-curves are derived from collecting *p*-values from simulations of *p*-hacked studies. *P*-hacking is simulated the same way as in Figure 1.