aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2014, Vol. 9(6) 666–681

DOI: 10.1177/1745691614553988
pps.sagepub.com
SAGE

p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results

Uri Simonsohn[1], Leif D. Nelson[2], and Joseph P. Simmons[1]
[1]University of Pennsylvania and [2]University of California, Berkeley

Abstract
Journals tend to publish only statistically significant evidence, creating a scientific record that markedly overstates the size of effects. We provide a new tool that corrects for this bias without requiring access to nonsignificant results. It capitalizes on the fact that the distribution of significant *p* values, *p*-curve, is a function of the true underlying effect. Researchers armed only with sample sizes and test results of the published findings can correct for publication bias. We validate the technique with simulations and by reanalyzing data from the Many-Labs Replication project. We demonstrate that *p*-curve can arrive at conclusions opposite that of existing tools by reanalyzing the meta-analysis of the "choice overload" literature.

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

Scientific inquiry is concerned not only with establishing whether an empirical relationship holds, but also with estimating the size of that relationship. For example, policy makers not only want to know whether a particular policy will produce the desired effect, but also whether the size of that effect is large enough to justify putting the policy into action. To estimate effect sizes of particular relationships, scientists often conduct meta-analyses, combining the results of many similar studies into a single effect size estimate. Unfortunately, because of biases in the publication process, producing an accurate effect size estimate is often extremely difficult.

Scientific journals usually do not publish results unless they are statistically significant (henceforth assumed to correspond to $p \leq .05$), a fact we will refer to as *publication bias* (see e.g., Fanelli, 2012; Rosenthal, 1979; Sterling, 1959). Because overestimated effect sizes are more likely to be significant than are underestimated ones, the published record systematically overestimates effect sizes (Hedges, 1984; Ioannidis, 2008; Lane & Dunlap, 1978).

To illustrate, imagine a researcher investigating whether people in a happy mood are willing to pay more for experiences than are people in a sad mood. She randomly assigns 40 people to watch either a happy video or a sad video and then measures their willingness to pay for a ticket to see their favorite band in concert. With 20 people per condition, the two condition means would have to differ by at least .64 standard deviations (i.e., $\hat{d} \geq .64$) for them to be significantly different.[1] Thus, no matter what the true effect size is, with 20 observations per condition, the average significant effect size must be at least .64 standard deviations. In fact, even if an effect does not exist at all ($d = 0$), the effect size estimated from just the significant studies will be large, with the means differing by $\hat{d} = .77$ standard deviations (see Fig. 2A).

Scientists wanting to estimate the true size of an effect need to correct the inflated effect size estimates that publication bias produces. In this article, we introduce a new and better method for doing so. This method derives effect size estimates from *p*-curve, the distribution of significant *p* values across a set of studies (Simonsohn, Nelson, & Simmons, 2014). We show that this simple technique, which requires only that we obtain significant *p* values from published studies, allows scientists to much more accurately estimate true effect sizes in the presence

Corresponding Author:
Uri Simonsohn, University of Pennsylvania - The Wharton School, 500 Huntsmann Hall, 3730 Walnut Street, Philadelphia, PA 19104
E-mail: uws@wharton.upenn.edu

of publication bias. When the publication process suppresses nonsignificant findings, *p*-curve's effect size estimates dramatically outperform those generated by the most commonly used technique for publication-bias correction.

p-Curve and Effect Size

p-curve is the distribution of statistically significant *p* values ($p < .05$) across a set of studies.[2] For example, if four studies report critical *p* values of .043, .039, .021, and .057, *p*-curve for this set of studies would include all of those that are below .05: .043, .039, .021, but not .057. In a previous article, we showed how *p*-curve's shape diagnoses whether a set of studies contains evidential value or not (Simonsohn et al., 2014). In this article, we show how one can use *p*-curve's shape to estimate the average true effect size across the set of studies included in *p*-curve. Note that in the face of publication bias, the average true effect size will differ from the average observed effect size, and that *p*-curve will estimate the former. Here is an intuitive way to think of *p*-curve's estimate: It is the average effect size one expects to get if one were to rerun all studies included in *p*-curve.

A *p* value reflects the likelihood of observing at least as extreme an estimate if there is truly no effect ($d = 0$). Thus, by definition, if an effect is not real, then 5% of *p* values will be below .05, 4% will be below .04, 3% will be below .03, 2% will be below .02, and 1% will be below .01. Thus, under conditions of no effect ($d = 0$), there will be as many *p* values between .04 and .05 as between .00 and .01, and *p*-curve's expected shape is *uniform*.

If an effect exists, then *p*-curve's shape changes. Its expected distribution will be right-skewed: We expect to observe more low significant *p* values ($p < .01$) than high significant *p* values ($.04 < p < .05$; Cumming, 2008; Hung, O'Neill, Bauer, & Kohne, 1997; Simonsohn et al., 2014; Wallis, 1942). For any given sample size, the bigger the effect, the more right-skewed the expected *p*-curve becomes. Figure 1 shows some examples.

Conveniently, for a particular statistical test, *p*-curve's expected shape is solely a function of sample size and effect size. Because of this, knowing *p*-curve's shape and the sample size for a set of studies allows one to compute the effect size. Holding sample size constant, a greater proportion of small significant *p* values (i.e., a more extreme right skew) implies a larger effect size.

To get an intuition for how to estimate effect sizes using *p*-curve, consider a difference-of-means test with $n = 20$ per cell. As shown in Figure 1, if the true effect size is $d = .42$, then 38% of significant *p* values are expected to be below .01. If the true effect size is $d = .91$, then 71% of significant *p* values are expected to be below .01. Thus, for a set of studies with $n = 20$ per sample, if 38% of significant *p* values are below .01, then our best guess of these studies' average effect size would be $\hat{d} = .42$. If 71% of significant *p* values are below .01, then our best guess of these studies' average effect size would be $\hat{d} = .91$.

More generally, for an observed set of significant results, one can identify the expected *p*-curve that most closely resembles the observed *p*-curve, and then identify the effect size estimate corresponding to that *p*-curve. Because the shape of *p*-curve is a function exclusively of sample size and effect size, and sample size is observed, we simply find the effect size $\hat{d}$ that obtains the best overall fit. In the Appendix, we provide a detailed account (and R code) of how this is done.[3]

Selectively Reporting Significant Studies

As described above, scientists interested in estimating effect sizes must correct for the fact that nonsignificant findings are much less likely to be published than are significant findings. This fact has long been recognized and a variety of corrective techniques have been proposed (for a review, see Rothstein, Sutton, & Borenstein, 2005).

The most common technique is known as *Trim and Fill* (Duval & Tweedie, 2000a, 2000b).[4] Although Trim and Fill is in common use, it rests on the unlikely assumption that the selective reporting of studies is driven by effect size rather than statistical significance.[5] That is, it assumes that the publication process suppresses the publication of small effects (regardless of significance) rather than nonsignificant results (regardless of effect size).[6]

However, publication bias in psychology (Rosenthal, 1979; Sterling, Rosenbaum, & Weinkam, 1995) and other fields (Ashenfelter, Harmon, & Oosterbeek, 1999; Gerber & Malhotra, 2008) primarily involves the suppression of nonsignificant results. As shown below, when the publication process suppresses nonsignificant findings, Trim and Fill is woefully inadequate as a corrective technique. *p*-curve performs much better.[7]

We conducted simulations to examine how well *p*-curve corrects for the selective reporting of statistically significant studies, and contrasted its performance with that of Trim and Fill. Specifically, we simulated studies testing a directional prediction with a two-sided difference-of-means test, pooled all the statistically significant studies with the predicted effect into a meta-analysis that included about 5,000 studies, and, to capture the selective reporting of significant studies, we estimated the true effect size based on the statistically significant studies alone.

We estimated the effect size using three different approaches. First, we conducted a traditional fixed-effect

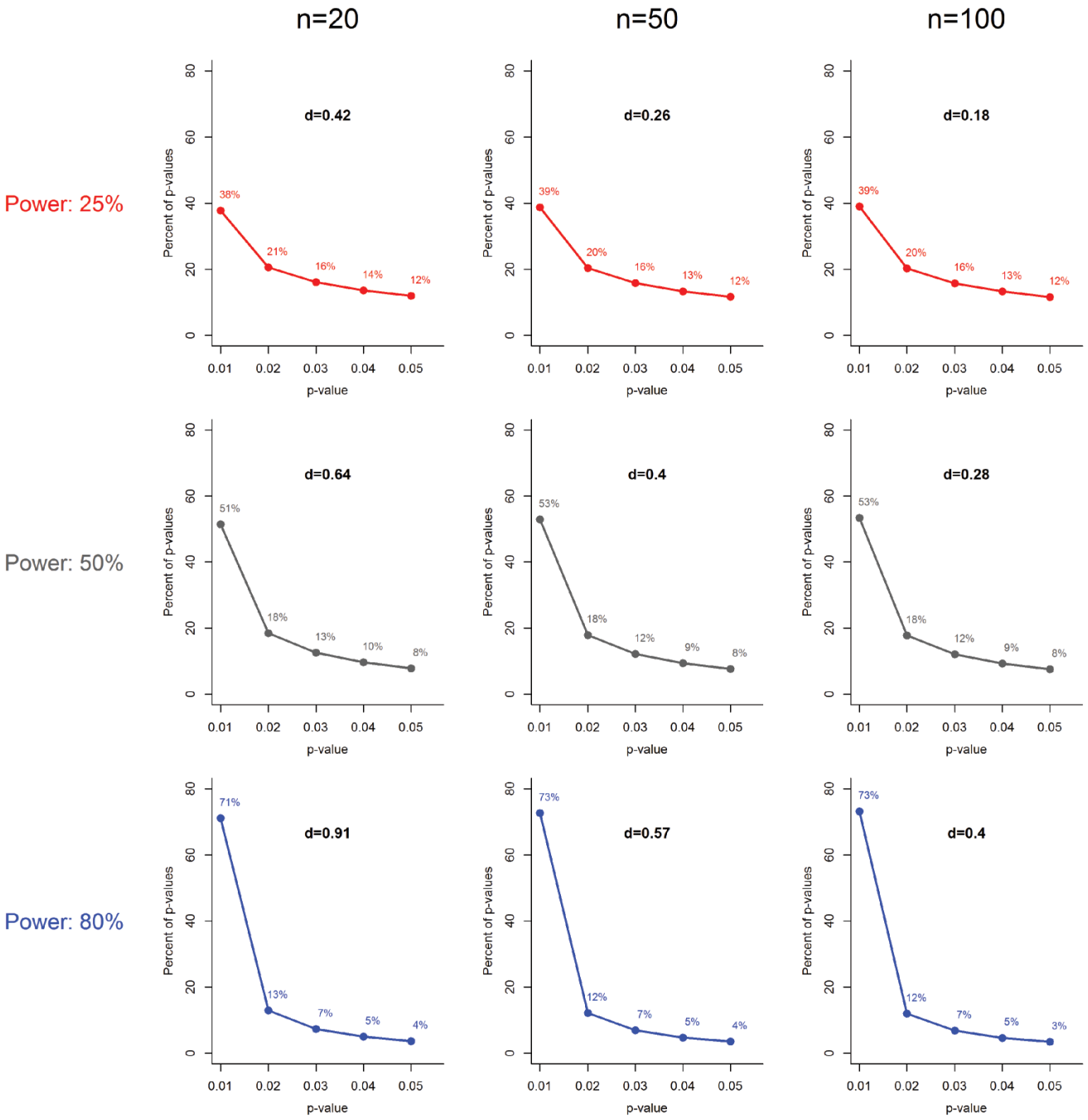

Fig. 1. *p*-curve's shape as a function of sample size and effect size. Expected *p*-curves for two-sample difference-of-means *t* tests, with *n* subjects per cell, where population means differ by *d* standard deviations from each other. Note that for a given level of power, *p*-curve is almost the same for every underlying sample-size and effect-size combination.[20] Plotted results are obtained from noncentral *t* distributions (see Appendix for details).

meta-analysis, computing the weighted average of observed effect sizes across the significant studies, without any correction for selectively reporting studies. Second, we corrected this estimate using the Trim and Fill procedure (Duval & Tweedie, 2000a, 2000b). Third, we estimated effect size using *p*-curve.

As shown in Figure 2, we conducted these simulations under a number of conditions. Panel A reports results assuming that all studies within a meta-analysis have the same sample size ($n = 20$ per cell) and effect size (shown on the *x*-axis). Panel B reports results allowing for studies within a meta-analysis to vary in sample size between

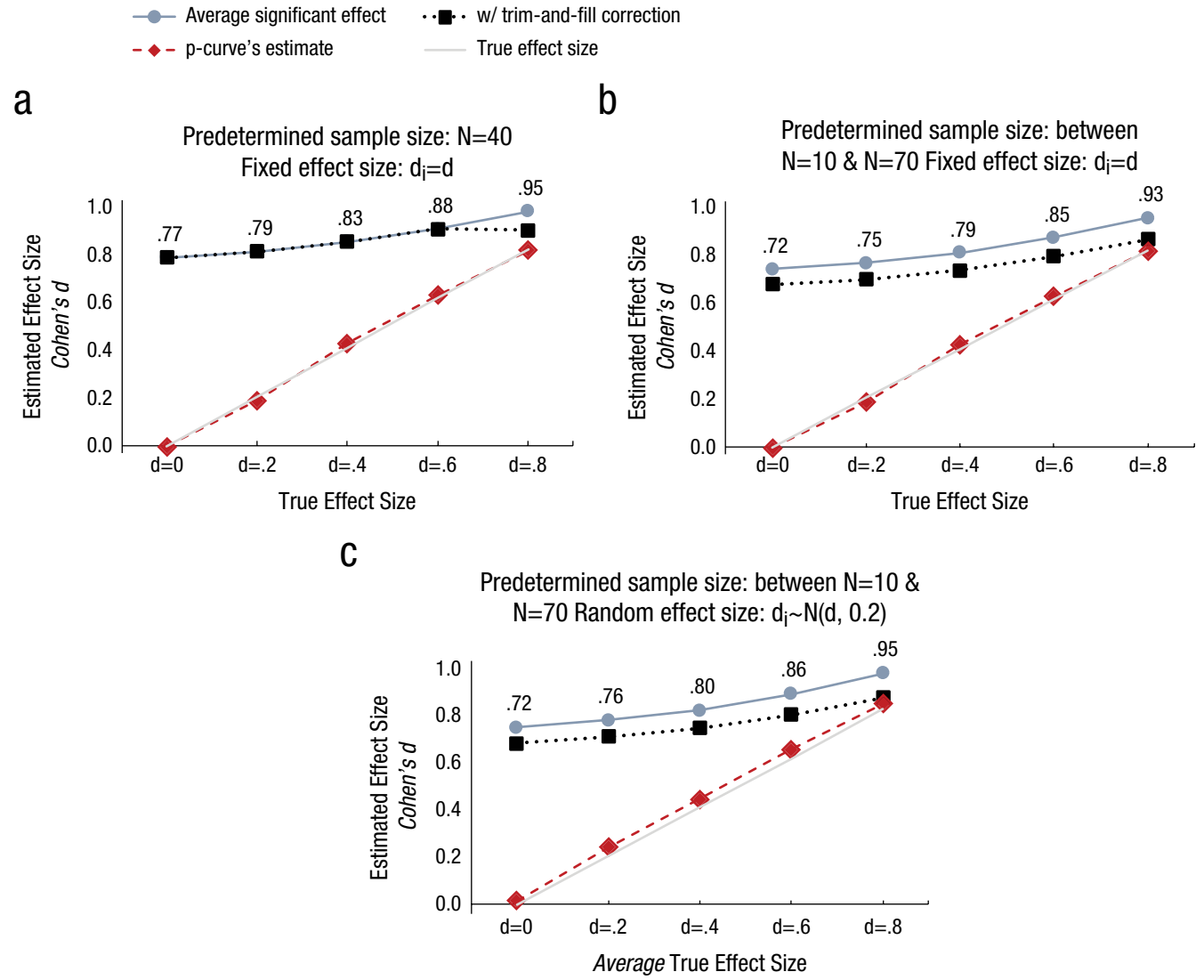

Fig. 2. Impact of selectively reporting significant studies. Each marker reports an effect size estimate (Cohen's *d*) based on a meta-analysis performed on about 5,000 statistically significant simulated studies. In Panel A, all studies have 20 observations per cell and assume a fixed effect size. In Panel B, the same number of statistically significant studies with each sample size between $n = 5$ and $n = 35$ per cell are included in the meta-analysis. Panel C is the same as Panel B, except effect sizes are drawn from a normal distribution with mean *d*, standard deviation $\sigma = .2$. Trim and Fill and *p*-curve estimates are based exclusively on $p < .05$ results.

$n = 5$ and $n = 35$ per cell. In Panel B's simulations, all studies within a meta-analysis had the same true effect size, and each meta-analysis included the same number of studies with each sample size.

The simulations in Panel C varied both sample size and true effect size across studies included in the same meta-analysis. For each simulated study, we first randomly drew a true effect size from a normal distribution with $\sigma = .2$ and the mean indicated on the graph's *x*-axis. We then drew observations from populations whose true means differed by that random effect while varying per-cell sample size to be between $n = 5$ and $n = 35$, and we pooled the set of statistically significant results. All studies within a meta-analysis had the same average true effect size, and each meta-analysis included the same number of statistical significant studies with each sample size.

Figure 2 displays the results, revealing several important facts. First, it shows the dramatic inflation of effect size generated by publication bias (Hedges, 1984; Ioannidis, 2008; Lane & Dunlap, 1978). The darker solid lines in the figure show that the average effect sizes of the subset of studies that are statistically significant are dramatically higher than the true effect sizes. Moreover, the estimates hardly vary as a function of true effect size. When samples are small, the subset of statistically

significant effect sizes contains almost no information about the true effect size.

Second, applying the Trim and Fill correction to these estimates does not make them meaningfully better. For example, when there was truly no effect ($d = 0$), Trim and Fill estimated the effect to be large, at least $d = .65$. When nonsignificant studies are not observed, the most popular corrective technique is not very corrective.

Third, in contrast to the other methods, *p*-curve fully corrects for the impact of selectively reporting studies. For example, in Panel A, we see that when the true effect size is $d = 0$, *p*-curve correctly estimates the effect to be zero, despite *p*-curve being based exclusively on observed estimates of $\hat{d} > .64$, with an average observed effect size of $\hat{d} = .77$. Panels B and C show that the accuracy of *p*-curve does not rely on homogeneity of sample size nor effect size. In all cases, *p*-curve is accurate and the other methods are not.[8]

The results from Figure 2 assess the performance of Trim and Fill when performed only on the subset of statistically significant findings, showing that it provides a minimal improvement over the naive average effect size. In Supplement 3, we show that adding nonsignificant findings to the set analyzed via Trim and Fill, even doubling the total number of studies, does not noticeably improve the corrective abilities of Trim and Fill.

p-Hacking

Researchers not only selectively report studies, they also selectively report analyses within a study (Cole, 1957; Simmons, Nelson, & Simonsohn, 2011). For example, a researcher may run a regression with and without outliers, with and without a covariate, with one and then another dependent variable, and then only report the significant analyses in the paper. We refer to this behavior as *p-hacking* (Simmons, Nelson, & Simonsohn, 2012; Simonsohn et al., 2014).

p-hacking enables researchers to find statistically significant results even when their samples are much too small to reliably detect the effect they are studying or even when they are studying an effect that is nonexistent. For this reason, existing methods for estimating effect sizes will be inflated, often dramatically, in the presence of *p*-hacking.

p-hacking biases *p*-curve's effect size estimates as well, but it does so in the opposite direction, leading one to underestimate effect sizes. To understand why, consider the effects that *p*-hacking has on *p*-curve's shape.

Because *p*-hacking leads researchers to quit conducting analyses upon obtaining a statistically significant finding, *p*-hacking is disproportionately likely to introduce "large" significant *p* values into the observed distribution (i.e., *p* values just below .05). As a result, *p*-hacking reduces the right skew of *p*-curve (Simonsohn et al., 2014). Because smaller effect sizes are associated with less right-skewed *p*-curves, *p*-hacking causes *p*-curve to underestimate effect sizes.

To explore the effects of *p*-hacking on effect size estimates, we simulated three common forms of *p*-hacking (John, Loewenstein, & Prelec, 2012; Simmons et al., 2011): achieving statistical significance by (a) data peeking (collecting more observations if an initial sample of observations does not obtain $p < .05$), (b) selectively reporting which of three dependent variables to report, and (c) selectively excluding outliers.

All of the simulations explored mean differences between two conditions starting with 20 observations in each. We varied the true effect size across simulations. For each simulation, we estimated effect size in three ways: (a) by computing the average of all effects (including all significant and nonsignificant findings), (b) by applying a Trim and Fill correction to only the statistically significant effects, and (c) by using *p*-curve. We report further details of these simulations in the next section; readers may choose to skip ahead to the Results section.[9]

Details

Figure 3A shows the results of simulations of data peeking. For each study, we first conducted a *t* test with $n = 20$ observations per cell. If the result was significant, we "published" it; if it was not, we added 10 observations to each sample, thus increasing the per-condition sample size from 20 to 30, and conducted a new *t* test. If this second result was significant, we "published" it; if not, it remained nonsignificant and hence "unpublished."

Figure 3B shows the results of simulations of selectively reporting among three dependent variables correlated with each other at $r = .5$. We conducted a *t* test on each of these dependent variables. Within a study, the first comparison to obtain significance was "published"; if all three *t* tests were nonsignificant, the study was "unpublished," and the analysis yielding the lowest *p* value was the one used to compute the average effect across all of the studies.

Figure 3C shows the results of simulations of selectively dropping outliers further than two standard deviations away from the sample mean. For each study, we conducted four *t* tests: one comparing both full samples, one dropping outliers from only the first sample, one dropping outliers from only the second sample, and one dropping outliers from both samples. Within a study, the first *t* test to obtain significance was "published." If all four *t* tests were nonsignificant, the study was "unpublished," and the analysis yielding the lowest *p* value was the one used to compute the average effect across all of the studies.

For each of these simulated forms of *p*-hacking we pooled all the significant results and estimated the underlying effect size using the three methods described above:

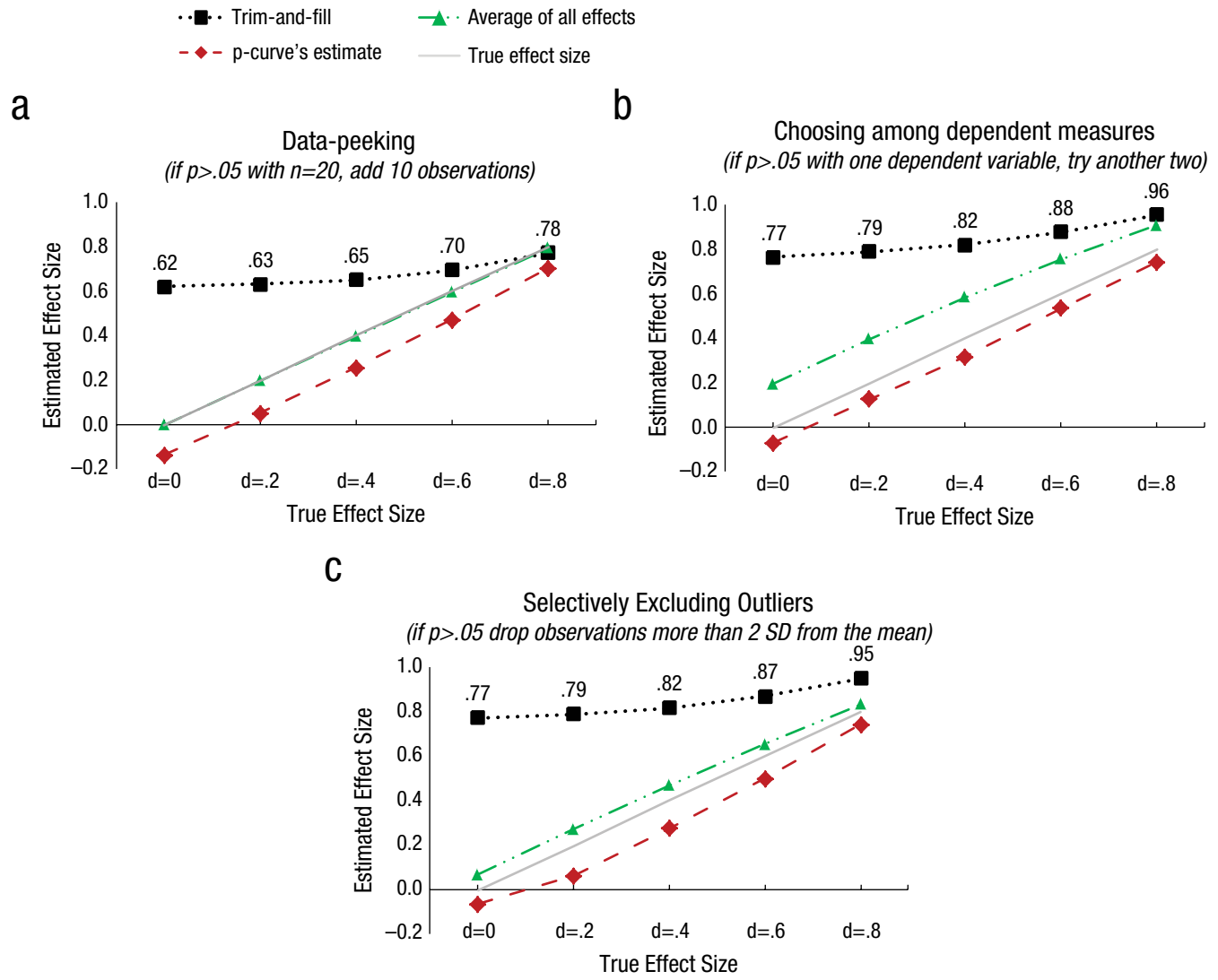

Fig. 3. Impact of *p*-hacking. Each marker reports an effect size estimate (Cohen's *d*) based on a meta-analysis performed on large numbers of studies. The triangle dash-dot line plots estimates based on all simulated studies (what would be estimated by a meta-analyst that obtained all studies ever conducted), and the square dotted line and diamond dashed line plot estimates based on the statistically significant subset. Each panel simulates a different form of *p*-hacking. A: Adding 10 observations per-cell if $p > .05$ is not achieved with $n = 20$. B: Analyzing three different dependent variables, reporting either the first to yield $p < .05$, or, if none are significant, the lowest *p* value obtained. C: Excluding observations further than two standard deviations from the condition's mean—first only from one condition, then the other, then both—if $p > .05$ has not yet been achieved. The square dotted lines are based on the last analysis conducted and hence may include bias from *p*-hacking; if they were based on the first, they would be free of *p*-hacking and hence always identical to the gray solid lines. Trim and Fill and *p*-curve estimates are based exclusively on $p < .05$ results.

the average of all studies regardless of significance, the estimate derived from applying the Trim and Fill correction to the significant studies, and *p*-curve's effect size estimates.

Results

Figure 3 shows the results. First, we again see that Trim and Fill does not adequately estimate effect sizes when the publication process suppresses nonsignificant results. Second, we see that these forms of *p*-hacking cause the *p*-curve to underestimate effect sizes. Interestingly, *p*-hacking biases not only the published record, but the totality of evidence, which includes all studies whether or not they are likely to be published. Thus, even if a meta-analyst were to gain access to every single study, eliminating publication bias via "brute force," *p*-hacking will cause effect sizes to be overestimated.

Thus, *p*-hacking will bias effect size estimates regardless of how effect sizes are estimated. Using traditional methods, *p*-hacking will bias effect sizes upwards; when using *p*-curve, *p*-hacking will bias effect sizes

downwards. Because the relative magnitude of these biases is situation specific, it is not possible to make general statements as to whether analyzing all studies ever conducted would outperform *p*-curve's estimate based only on the statistically significant subset.

Note that the results from Figure 3 assume that unpublished results remained *p* hacked (e.g., that observations that were dropped by a researcher but did not succeed in lowering the *p* value to $p < .05$, would also be excluded from the dataset given to the meta-analyst). If unpublished results were free of *p*-hacking—for example, if any dropped observations were reintroduced before the data were meta-analyzed—then the bias present in the meta-analysis of all conducted results would be reduced. It would still not be eliminated, because the published studies included in the meta-analysis are still biased upwards by *p*-hacking.

Precision

We have so far reported results from simulations involving large numbers of studies. We now turn to the issue of how much precision we may expect from *p*-curve's estimates when relying upon smaller sets of studies. Figure 4 reports results from simulations that varied true effect sizes, the number of studies included in *p*-curve, and the sample sizes of those studies (with per-cell sample size of either 20 or 50). The markers indicate the median effect size estimate across simulations, and the vertical bars indicate one standard error above and below that median (i.e., the standard deviation of that estimate across simulations).

As one may expect, *p*-curve is more precise when it is based on studies with more observations and when it is based on more studies. Less obvious, perhaps, is the fact that larger true effects also lead to more precision. This occurs because *p*-curve's expected shape quickly becomes very right-skewed as effect size increases, reducing the variance in skew of observed *p*-curves.

Demonstrations

In this section, we provide two demonstrations. The first relies on data from the "Many-Labs replication project" (Klein et al., 2014), in which 36 different labs around the world collaborated to run the exact same set of studies and reported all results regardless of statistical significance. This provides a unique opportunity to assess the performance of *p*-curve in a realistic environment—using real studies, real dependent variables, and real participants—where we nevertheless observe all studies conducted, regardless of outcome.

The second demonstration revisits the meta-analysis of the popular psychology literature on choice overload (Scheibehenne, Greifendeder, & Todd, 2010). This example demonstrates a situation in which *p*-curve and traditional meta-analytical tools arrive at different answers, suggesting different paths for what future empirical work on the topic ought to seek to accomplish.

Demonstration 1. Many Labs Replication Project

Klein et al. (2014) conducted replications examining 13 different "effects" across 36 labs (data available from https://osf.io/wx7ck/). We can use these data to assess how well *p*-curve corrects for publication bias in a realistic setting by comparing the effect size estimate we obtain from *p* curving only the subset of significant results to that obtained by averaging the results from all labs.

For this assessment of performance to make sense, and for the aggregate average to be a valid proxy for truth, we need to believe that the studies that worked and did not work were examining the same average effect—that they differed only because of sampling error. Otherwise, if *p*-curve recovers one estimate, and the aggregate average another, we don't know if *p*-curve performed poorly, or if it is correctly indicating that the significant and nonsignificant studies were investigating a different underlying effect. We thus focus on effects that proved homogeneous across the different labs.[10]

The two most homogenous effects were the sunk costs fallacy (as studied by Oppenheimer, Meyvis, & Davidenko, 2009) and the Asian disease problem (Tversky & Kahneman, 1981).[11] Conveniently, these two effects were associated with very different replication rates. Only 50% of labs obtained a significant result for the sunk cost fallacy, the lowest in the set of effects deemed "replicable," whereas 86% of the studies investigating the Asian disease problem were significant.[12]

Figure 5A shows the resulting *p*-curves. Both are right skewed, but Asian disease's *p*-curve was more so. Whereas 83% of the Asian Disease Problem's significant *p* values were below .01, only 31% of the Sunk Cost Fallacy's significant *p* values were below .01. Figure 5B reports the resulting effect size estimates, comparing *p*-curve's estimates to a naive estimate, computed by averaging the effect size observed across the significant studies, and an earnest estimate, computed by averaging the effect size across all studies, regardless of significance. Because these results were not *p* hacked, we can safely assume that the earnest estimate represents the best estimate of the true effect size.[13]

The bias of the naive estimate is small for the Asian disease problem, as a large proportion of those studies were significant. It estimates a true effect size of .66, whereas the average across all studies was .60. Reassuringly, *p*-curve's estimate agrees with the earnest

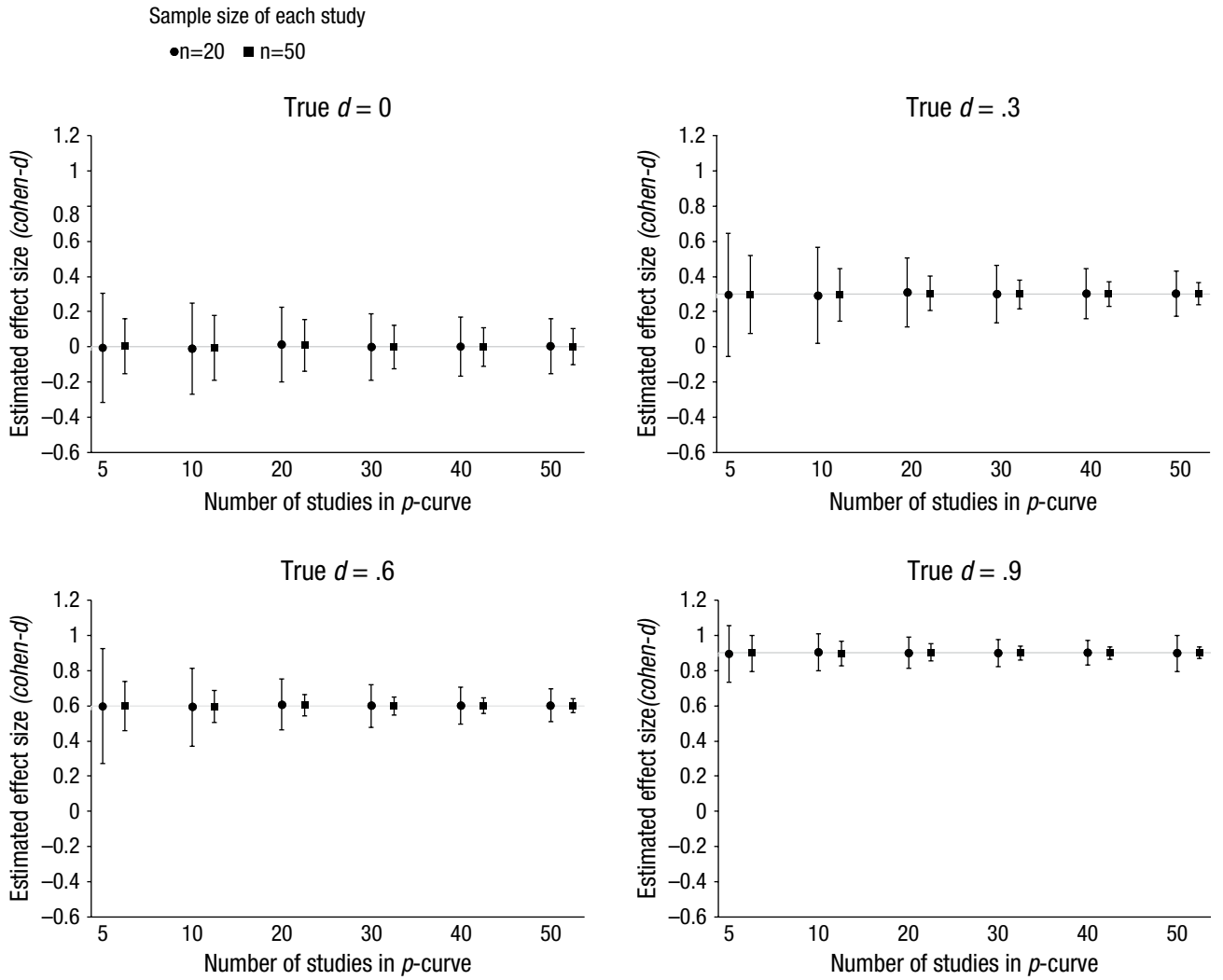

Fig. 4. Precision of *p*-curve. Each marker reports the median effect size obtained across 1,000 simulations of meta-analyses including between 5 and 50 studies, each with the same sample size (n = 20/50 per cell) and underlying true effect. Vertical bars show one standard error above and one standard error below the estimate. Standard errors are computed as the standard deviation of the parameter estimate across simulations.

estimate, and thus corrects little when little needs to be corrected. The bias of the naive estimate for the sunk cost fallacy is much larger, estimating a true effect size of .46 when the average across all studies was .31. Reassuringly, *p*-curve's estimate again agrees with the earnest estimate and thus corrects more when more needs to be corrected.[14]

Demonstration 2. Choice overload

The choice overload literature in psychology examines whether an increase in the number of options available leads to negative consequences, such as a decrease in motivation to choose or satisfaction with the option that is ultimately chosen. Scheibehenne et al. (2010) conducted a meta-analysis combining published and unpublished studies, obtaining an overall mean effect size of "virtually zero" (p. 409). This result implies that changes in choice-set size are inconsequential.

As Chernev, Bockenholt, and Goodman (2010) commented:

> "… studies often include two conditions: one designed to show that the effect of the construct (e.g., choice overload) is present, *and another one designed to document the directionally opposite* (e.g., more-is-better) effect […] combining their effect sizes to test their average effect leads to a potentially biased interpretation of the underlying effects." (p. 427, emphasis added)

In line with their (correct, in our view) observation, we split the studies analyzed by Scheibehenne et al. into two groups: one showing "choice is good" (positive

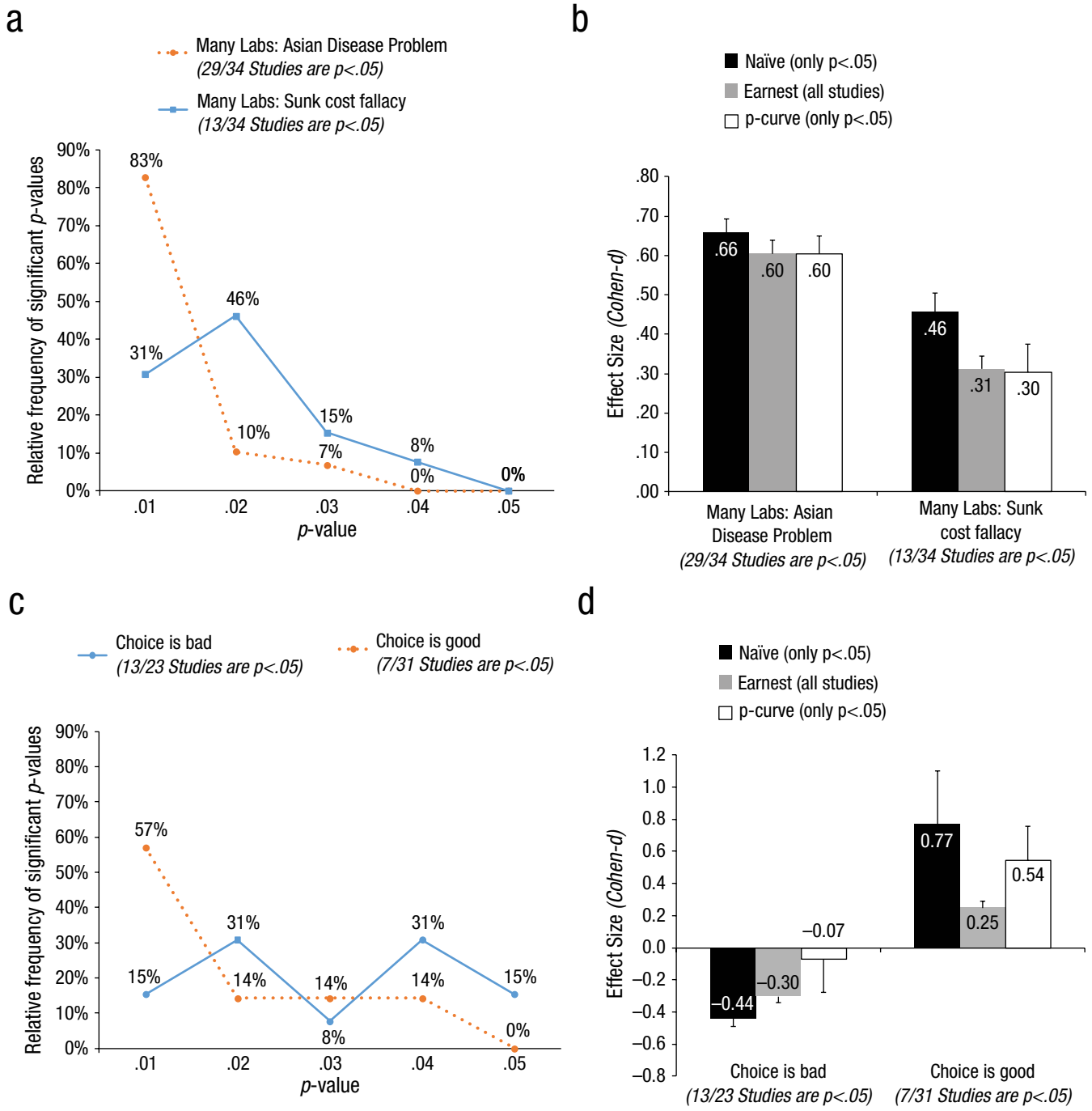

Fig. 5. Demonstrations of *p*-curve. Panel A depicts *p*-curves for studies reported in the Many Labs replication project (Klein et al., 2014), and Panel C depicts *p*-curves for studies included in the meta-analysis of the impact of choice-set size on consumer outcomes (Scheibehenne, Greifendeder, & Todd, 2010). The latter are split into studies showing a positive effect and a negative effect, as suggested by a commentary on that meta-analysis (Chernev, Bockenholt, & Goodman, 2010). Panels B and D depict the corresponding effect size estimates obtained via traditional meta-analytical tools applied only to statistically significant studies (naive), to all studies available (earnest), and from applying *p*-curve to the significant subset. Vertical bars correspond to one standard error.

coefficient of choice set size), and the other showing "choice is bad" (negative coefficient). We estimated effect sizes separately for both sets of studies, effectively asking two conceptually and statistically independent questions:

(1) In studies showing that more choice is good, how good is choice?
(2) In studies showing that more choice is bad, how bad is choice?

The results are reported in Figure 5D. When we limit our estimation to statistically significant studies and average across them, both effect size estimates are sizeable; both get closer to zero when we add the nonsignificant studies. Interestingly, when we apply *p*-curves to the significant findings, the choice-is-bad effect gets smaller, becoming effectively zero, whereas the choice-is-good effect gets larger. The right skewed *p*-curve for choice-is-good and flat *p*-curve for choice-is-bad, depicted in Figure 5C, reveal why *p*-curve estimates move in opposite directions for the two sets of studies.

Some of the error bars in Figure 5D are large. Particularly relevant is the standard error for the $\hat{d} = -.07$ estimate for choice-is-bad. *P*-curve's estimate is not conclusively saying the effect is 0; it is saying the best guess is close to zero, but that the data are consistent with practically and theoretically relevant effect sizes (of both signs) also.

Our interpretation of Figure 5D is the following. If we conduct traditional meta-analysis on the available evidence, we empirically verify the concern expressed by Chernev et al. (2010): under predictable circumstances choice is good, and under predictable circumstances choice is bad. This would suggest that future work in the choice overload literature could safely and constructively consist of conceptual replications that seek to learn more about moderators and mediators of the choice-is-bad effect.

If we conduct meta-analysis based on *p*-curve, the conclusion is different. Yes, the evidence does support the (uninteresting) notion that under predictable circumstances choice is good, but it neither confirms nor denies the (much more interesting) notion that under predictable circumstances choice is bad. The conclusion based on *p*-curve suggests that future work should focus on examining if the basic phenomenon of choice-is-bad can indeed be reliably obtained. Properly powered direct replications of the original demonstrations are something top journals may be less inclined to publish based on the traditional meta-analysis result ("We already know this!") but more inclined to publish based on our *p*-curve results ("We don't really know the answer to this yet"). Thus, given the same data, *p*-curve and traditional meta-analysis can suggest very different paths forward.

Limitations

Limitation 1. Simple effects from attenuated interactions cannot be analyzed

The validity of *p*-curve rests on the assumption that a lower *p* value does not increase the likelihood of publication once it crosses the significance threshold. For example, a significance criterion of .05 assumes that a *p* value of .008 would have been publishable if it had instead been .038.

Though this assumption usually holds, studies investigating attenuated interactions often violate it. An attenuated interaction hypothesis is one that predicts that an effect will be smaller under one condition than under a different condition. For example, the hypothesis that the effect of gender on height will be smaller for children than for adults is an attenuated interaction hypothesis. In this example, the *unattenuated simple effect* is the (larger) effect of gender on height for adults, whereas the *attenuated simple effect* is the effect of gender on height for children.

Researchers interested in publishing attenuated interactions need the interaction terms to be significant (Gelman & Stern, 2006). But for the interaction term to be significant, the (larger) unattenuated simple effect needs to have an even lower *p* value. For example, if the simple effect of gender on adults' heights was associated with $p = .038$, it is unlikely that the interaction would be significant. This means that the de facto significance criterion for unattenuated simple effects in this design is smaller than for the other *p* values. As a result, the inclusion of *p* values for unattenuated simple effects will result in a *p*-curve that overestimates effect sizes. The inclusion of only the interaction term *p* value would leave *p*-curve unbiased. As a result, for studies hypothesizing attenuated interactions, we recommend never including results from simple effects in *p*-curve.

It is worth noting that this problem does not apply to studies predicting reversing interactions, involving an effect being observed under one condition but then the opposite effect being observed under a different condition. In this case, both simple effects may be included in *p*-curve.

Limitation 2. p-*curve ignores nonsignificant results*

Another limitation is that *p*-curve ignores information from nonsignificant studies ($p > .05$). Because this limits the sample size of studies under consideration, *p*-curve is more likely to provide a noisy effect size estimate. This is a necessary limitation of *p*-curve: Because we do not know what publication pressures occur above .05, we do not know what the distribution of *p* values should be above .05. Note that although excluding nonsignificant results makes *p*-curve noisier (less efficient), it does not make *p*-curve biased.

Limitation 3. Downward bias with p-*hacking*

As we discussed in some detail when presenting the results from Figure 3, *p*-hacking can bias effect size

estimates from *p*-curve downwards. In our simulations, however, this bias is mild enough to be ignorable.

Limitation 4. Moderation within meta-analysis

In its current form *p*-curve estimates a single average for all studies included in it. To examine if a variable moderates the effect size of interest across studies, then separate *p*-curves would need to be estimated. For example, to examine if anchoring studies run in the lab have different effect sizes than anchoring studies run outside the lab, one would perform one *p*-curve for lab studies and one *p*-curve for nonlab studies, rather than a single *p*-curve that includes a moderator variable. It seems likely that *p*-curve can be modified to incorporate moderation within a single analysis, but we have not explored that possibility.

Another Use of *p*-curve: Average Power

p-curve's shapes is closely tied to statistical power, the probability that a study obtains a significant result. For a given statistical test, both power and *p*-curve depend exclusively on the size of the sample and the size of the effect. This means that *p*-curve can be used to estimate the average underlying statistical power of a set of studies. As with effect sizes, *p*-curve's estimate of power will correct for the inflated estimates that arise from the privileged publication of significant results.

Estimating the publication-bias corrected estimate of the average power of a set of studies can be useful for at least two purposes. First, many scientists are intrinsically interested in assessing the statistical power of published research (see e.g., Button et al., 2013; Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1989). But to carry out their calculations they have either (a) relied on arbitrary effect size assumptions (e.g., small, medium, and large) and asked how much power do the observed sample sizes have to detect effects of that size, or (b) computed the average observed effect size (which is inflated by publication bias and causes problems if effect size is heterogeneous) and computed the post-hoc power for that effect. With *p*-curve, those arbitrary assumptions are no longer needed, and we can estimate the actual underlying power correcting for publication bias.

Second, published results often provide insufficient details to compute effect size. For example, effect size in mixed between-/within-subject designs depends on within-participant correlations across observations—a metric that is often not reported. Nevertheless, the true underlying power of the reported test statistic (e.g., *F* test in an analysis of variance) and hence the resulting *p*-curve is of course influenced by that parameter whether it is reported or not and, hence, so is the average power estimated via *p*-curve. When aggregating results reported in insufficient detail, one may estimate average underlying power and then convert the result into an intuitive metric of effect size taking into account the underlying sample size.

User Guide

Upon identifying the effect of interest (e.g., the impact of a high vs. low anchor value on monetary valuations) and the selection criterion for studies to include (e.g., all studies that cite Chapman and Johnson, 1999, and use a monetary dependent variable), the researcher must identify, for each study, the test statistic associated with testing the null that the effect of interest is zero. The set of all such tests is then submitted to a *p*-curve analysis, either using the R Code included in this article or the online web-app available at www.p-curve.com.

p-curve assumes these tests are statistically independent from one another. If multiple results from the same participants are reported in a paper, only one of them may be included in a given *p*-curve. If only one of the results is pertinent to the meta-analysis then the decision is easy: Only the pertinent test is selected.

For example, if an anchoring study presented participants with both a high and a low anchor and then asked them to separately value both a coffee mug and a pen, then a meta-analyst only concerned with the impact of anchors on mugs would only include the mug result.

If multiple results are pertinent to the meta-analyst's question, then a decision must be made. For example, if both the pen and mug were relevant to the question of interest to the meta-analysis, then *p*-curve may include either the test statistic associated with an analysis of the mug, with an analysis of the pen, or with an analysis of a composite of the mug and the pen (e.g., the averaged or summed ratings of the mug and pen). But *p*-curve can never include results from more than one of these analyses. Importantly, *p*-curve should also never include the average *p* value of multiple tests. For example, if the anchoring effect for coffee mugs was $t(38) = 2.12$, $p = .04$ and the anchoring effect for pens was $t(38) = 2.43$, $p = .02$, *p*-curve should not include the average *p* value, $p = .03$, nor the *p* value associated with the average test statistic, $p = .028$. It should include either $p = .02$ or $p = .04$.[15,16]

When choices among multiple tests must be made, we recommend adhering to a prespecified selection rule (e.g., the test reported first) and then computing and reporting the result obtained under a different rule (e.g., the test reported last).

As implemented here, all test results entered into *p*-curve are assumed to be either examining effects of the

same sign (e.g., all significant effects in *p*-curve show that anchoring occurs, and none show that the reverse of anchoring occurs) or that the sign of the effect is not relevant (e.g., when computing average statistical power across heterogeneous findings). If neither of these conditions is met, two separate *p*-curves should be conducted, one for positive effects and one for negative ones.[17]

For scientific results to be interpretable, it is imperative that researchers disclose how they resolved ambiguities surrounding the collection and analysis of data (Simmons et al., 2011). For *p*-curve users in particular, this is easily achieved by supplementing their explicit identification of a study selection rule with a *p*-curve disclosure table (Simonsohn et al., 2014).

To examine the impact of a discrete moderator on effect size, the simplest way to proceed is to split up the studies into different subgroups (e.g., studies performed in the United States make up one subgroup, studies performed outside the United States make up another) and then *p*-curve is applied separately to each group, obtaining a separate effect size estimate for each subgroup. In its current implementation, *p*-curve does not allow for the analysis of continuous moderators (see the Limitations section).

Overview of Supplementary Materials

The Supplementary materials include the following sections.

1. Robustness of *p*-curve to data that are not normally distributed
2. Robustness of *p*-curve to heterogeneity of effect size
3. Trim-and-Fill performance when some p>.05 are observed
4. Alternative loss functions (to the one from the appendix)
5. R-Code for every result in this article (also available here: http://www.p-curve.com/Supplement/Rcode)

Conclusions

The selective publication of significant studies and analyses leads the published record to overestimate the size of effects. We have shown that one can use the distribution of significant *p* values, *p*-curve, to easily and effectively estimate effect sizes that correct for the selective reporting of studies, vastly outperforming the most commonly used alternative, Trim and Fill. *p*-curve also outperforms existing methods when researchers selectively report analyses—when they *p* hack—but for many forms of *p*-hacking it underestimates true effect sizes. However, the presence of *p*-hacking biases all known methods of effect size estimation, even when one averages across every study ever conducted. Overall, *p*-curve seems to be the best tool for estimating effect size when the publication process predicts statistically significant results.

Technical Appendix: Approach and Algorithm for Estimating Effect Size Using *p*-curve

The goal is to determine the underlying effect size that leads to an expected *p*-curve that best fits the observed *p*-curve. This requires three steps: (a) linking underlying effect size with expected *p*-curves; (b) defining a *loss function*, a metric of how well a given expected *p*-curve fits the observed *p*-curve; and (c) finding the effect size that minimizes that loss function.[18]

A1. Linking effect size with expected p-curve

We assume here familiarity with noncentral distributions. For most readers that's a terrible assumption, but it can be remedied by consulting Supplement 1 in Simonsohn et al. (2014).

For simplicity, we focus on independent two-sample difference of means *t* tests performed on samples of the same size (*n*). As an introduction, let's go over how we constructed Figure 1. The top left panel shows the expected *p*-curve when $n = 20$ and $d = .4164$ (shown as .42). To obtain that expected *p*-curve, we first identify the critical *t* values that lead to $p = .01,.02\ldots.05$ with a degree of freedom of 38.

Using R syntax, this involves:

```
x5 = qt(.975,df = 38)
x4 = qt(.98, df = 38)
x3 = qt(.985,df = 38)
x2 = qt(.99, df = 38)
x1 = qt(.995,df = 38)
```

For example, x5 = 2.0244, which means that $t(38) = 2.0244$ leads exactly to $p = .05$ (for a two-sided test).

Next, we rely on the noncentral student distribution to see how likely one is to obtain *t* values more extreme than each of x1, x2…x5. For the parameters above, $n = 20$, $\delta = .4164$, This involves using the noncentrality parameter ncp $= \sqrt{\frac{n}{2}}\,\delta = \sqrt{\frac{20}{2}}\,.4164$

Starting with x5 (again, the lowest *t* value that leads to a statistically significant result), we compute, in R-syntax again)

```
1-pt(x5,df=38,ncp=sqrt(20/2)*.4164) =.25
```

That is the probability of obtaining $t \geq 2.0244$, and hence $p \leq .05$ when $n = 20$ and $\delta = .042$—it is, hence, the statistical power of the test. There is a 25% chance of obtaining a statistically significant result with a study of those characteristics. The top-left panel of Figure 1 identifies what share of 25% of tests will be $p < .01$, $.01 < p < .02$, etc. Let's determine what share of 25% of tests will be $p < .01$:

```
1-pt(x1,df=38,ncp=sqrt(20/2)*.4164)
```

We get: .094

This means that with $n = 20$, there is a 9.4% chance of obtaining $p < .01$ when $\delta = .42$.

Among all attempted studies, 25% are $p < .05$ and 9.4% are $p < .01$. Therefore, the share of significant results that are $p < .01$ is 38% (9.4/25; see Fig. 1). Proceeding analogously with x2, x3, and x4, we obtain the rest of the plotted numbers: the histogram version of *p*-curve.

For estimating effect size we treat *p*-curve as the continuous distribution that it is. We proceed analogously, but we do not limit ourselves to the five discrete points x1–x5; instead, we create a function that maps every possible statistically significant *t* value to the probability of obtaining a *t* value at least as large. This is effectively the *p* value of the *p* value, which we referred to as the pp-*value* (Simonsohn et al., 2014).

For *t* value *i*, t_i, the probability of observing at least as large a significant *t* value is:

$$pp(ti) = prob(t > t_i \mid df,\ ncp, p < .05)$$

It is useful to get $p < .05$ out of the conditional. Let's define,[19]

$$power = prob(p < .05 \mid df, ncp).$$

Then

$$pp(t_i) = (prob(t > t_i \mid df, ncp) - power) / power$$

$pp(t_i)$, then, is the *cumulative distribution function* (*c.d.f.*) of *p*-curve, a function mapping sample and effect size onto expected *p*-curve. Because we observe sample size, it effectively maps effect size onto expected *p*-curve.

A2. Defining a loss function

For every candidate effect size d_i, then, there is a $pp(t \mid d_i)$ function that gives every possible *t* value a probability of observing at least as extreme a value. When $d_i = \delta$ (when the candidate effect size equals the true effect size), *pp* values will be distributed uniformly for reasons entirely analogous to why *p* values are distributed uniformly under the null (which we explain in the main text).

In light of this, we define how well a given candidate effect size d_i fits the data—how well the expected *p*-curve fits the observed *p*-curve—by assessing how close to a uniform distribution the set of observed *pp* values are. Many techniques exist to compare empirical with expected distributions—we rely on the robust and simple Kolmogorov-Smirnov (KS) statistic, which computes the maximum observed gap between the two *c.d.f.*s. The biggest gap, often represented by *D*, has a known asymptotic distribution that is used to convert the test into the KS-test *p* value, but we do not need this additional step. We simply use *D* as the metric of fit; as the *D* value increases, less of the observed *p*-curve is captured by the expected *p*-curve. *D* has an intuitive representation, if $D = .4$, then the biggest observed gap in (cumulative) *p*-curve is 40%. For example, 78% of *p* values are supposed to be $p < .041$, but only 38% of them are: 78% − 38% = 40%. In Supplement 4, we discuss alternatives to the KS test for measuring fit.

A3. Minimizing the loss

The last step consists of finding the candidate effect size that minimizes *D*. We rely on R's *optimize()* command for this, but we first exhaustively search the plausible space of effect size so as to (a) reduce the odds that *optimize()* lands on a local rather than global minimum and (b) provide a diagnostic plot of how well different effect size fit the observed *p*-curve. See R code below.

```
# R-CODE for estimating effect size via p-curve – written by Uri Simonsohn
#Define the loss function
loss=function(t_obs,df_obs,d_est) {
  t_obs=abs(t_obs)                          #Take absolute value of t-value (p-curve assumes same sign and/or sign does not matter)
  p_obs=2*(1-pt(t_obs,df=df_obs))           #Compute p-values of each t in t_obs so as to keep only p<.05 results
  t.sig=subset(t_obs,p_obs<.05)             #Significant t-values
  df.sig=subset(df_obs,p_obs<.05)           #d.f. associated with significant t.values
  ncp_est=sqrt((df.sig+2)/4)*d_est          #Compute noncentrality parameter for that sample size and candidate effect size
  tc=qt(.975,df.sig)                        #Compute critical t-value to get p=.05
  power_est=1-pt(tc,df.sig,ncp_est)         #Compute power for obtaining that t-value or bigger, given the noncentrality parameter
```

```
  p_larger=pt(t.sig,df=df.sig,ncp=ncp_est)    #Probability of obtaining a t-value bigger than the one that is observed (this is a vector)
  ppr=(p_larger-(1-power_est))/power_est      #Conditional probability of larger t-value given that it is p<.05, pp-values
  KSD=ks.test(ppr,punif)$statistic            #Kolmogorov Smirnov test on that vector against the theoretical U[0,1] distribution
  return(KSD) }

#Find the best fitting effect size (this also generates a diagnostic plot)

plotloss=function(t_obs,df_obs,dmin,dmax)  #Syntax, same as above plus: dmin/dmax: smallest/biggest d considered,
  { loss.all=c()                           #Vector where results of fit for each candidate effect size are stored
    di=c()                                 #Vector where the respective effect sizes are stored
    for (i in 0:((dmax-dmin)*100)) {       #Do a loop considering every effect size between dmin and dmax in steps of .01
      d=dmin+i/100                         #What effect size are we considering?
      di=c(di,d)                           #Add it to the vector of effect sizes
      options(warn=-1)                     #turn off warning because R often generates warnings when using noncentral pt() and
qt() that are inconsequential (they involve lack of precision at a degree where precision lacks practical relevance)
  loss.all=c(loss.all,loss(df_obs=df_obs,t_obs=t_obs,d_est=d)) #add loss for that effect size to the vector with all losses
  options(warn=0)                          #turn warnings back on
    }
  imin=match(min(loss.all),loss.all)       #Find the attempted effect size that leads to smallest los overall
  dstart=dmin+imin/100                     #Counting from dmin, what effect size is that?
  dhat=optimize(loss,c(dstart-.1,
  dstart+.1), df_obs=df_obs,t_obs=t_obs)   #Now optimize in the neighborhood of that effect size

#PLOT RESULTS

  plot(di,loss.all,xlab="Effect size\nCohen-d", ylab="Loss (D stat in KS test)",ylim=c(0,1), main="How well does each effect size fit?
  (lower is better)")
  points(dhat$minimum,dhat$objective,pch=19,col="red",cex=2)      #Put a red dot in the estimated effect size
#Add a label
  text(dhat$minimum,dhat$objective-.08,paste0("p-curve's estimate of effect size:\nd=",round(dhat$minimum,3)),col="red")
  return(dhat$minimum)
    }
#Example
  t_obs= c(1.7, 2.8, -3.1, 2.4)            # include one p > .05 and one negative t-value to highlight how we treat those
  df_obs=c(44, 75, 125, 200)
  plotloss(t_obs=t_obs,df_obs=df_obs,dmin=-1,dmax=1)
```


Acknowledgments

We received useful feedback for this project at the following seminar series and conferences: Wharton Decision Processes (March, 2012), Harvard NOM (April 2012), Columbia Marketing (May 2012), UCLA Marketing (May, 2012), BEAM – Cornell (June, 2012), Columbia Statistics (September 2012), Rutgers Psychology (October, 2012), Berkeley Initiative for Transparency in Social Science (December, 2012), UCSD Rady School (January, 2013), SPSP conference, New Orleans (January 2013), University of Southern California Marketing (February, 2013), the Solid Psychological Science Symposium, Nijmegen (June, 2013), and INSEAD (April, 2014). Any errors are our responsibility. This manuscript previously circulated as: Nelson, Simonsohn, & Simmons "P-Curve Fixes Publication Bias: Obtaining Unbiased Effect Size Estimates from Published Studies Alone."

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Supplemental Material

Additional supporting information may be found at http://pps.sagepub.com/content/by/supplemental-data or at http://www.p-curve.com/Supplement/Suplement_pcurve2.pdf

Notes

1. The two sample t test with 38 degrees of freedom obtains p = .05 if $t(38) = 2.024$. We can find the corresponding effect size for $t = 2.024$ by recalling that $t = \frac{M_2 - M_1}{SD\sqrt{2/n}}$ and $\hat{d} = \frac{M_2 - M_1}{SD}$, so $\hat{d} = t\sqrt{2/n}$. Needing $t \geq 2.024$ for statistical significance is hence mathematically equivalent to needing $\hat{d} \geq 2.024\sqrt{2/20} = .64$.
2. See the User Guide section for more details regarding what p values may and may not be included in p-curve.
3. For ease of exposition, throughout the article we focus on differences of means t tests and, hence, the t distribution.
4. Our assessment of Trim and Fill being the most commonly used corrective technique is informed by two sources. First, our casual observation indicates that, in psychology, review articles that correct for publication bias do so exclusively using Trim and Fill. Second, as of May 2014, Google Scholar indexes the

original Trim and Fill article as having 1,504 citations. Articles introducing other correction tools reviewed by Hedges and Vevea (2005) have about 50–150 citations. Thus, a rough estimate is that Trim and Fill is at least 10 times as popular as its competitors.

5. With this method, a meta-analyst explores the relation between the sample size and effect size of a set of studies, looking to see whether some studies appear to be missing as a result of publication bias. For example, if a very large-sample study estimates an effect size to be $\hat{d} = .40$, and most of the smaller studies estimate an effect to be greater than $\hat{d} = .40$, then it is presumed that smaller studies estimating effects less than $\hat{d} = .40$ were unpublished. Trim and Fill is an algorithm that trims (i.e., eliminates) some real studies, and fills in (i.e., introduces) some non-real ones, seeking to obtain a final set of studies in which there would be a similar number of small sample studies above and below $\hat{d} = .40$

6. Duval and Tweedie (2000a) write "Our key assumption is that the suppression has taken place in such a way that the […] most extreme negative values have been suppressed" (p. 91).

7. Hedges and Vevea (2005) reviewed additional publication bias correction tools that, as we do here, assume selective reporting based on *p* values. The approach most similar to ours is by Hedges (1984). Two key differences are that his approach does not eliminate bias due to selective reporting of studies (see his Fig. 4), and cannot be applied when all effects are of the same sign (Hedges & Vevea, 2005, p. 152).

8. To provide a more intuitive treatment of heterogeneity than that captured in Figure 2C's simulations, we performed additional simulations where half the studies had one true underlying effect size (d_1), and the other half had a different true underlying effect size (d_2). We then applied *p*-curve to that pooled set of statistically significant findings and verified that the estimated effect size, $\hat{d}$, corresponded to the average of the two true effect sizes. For example, if we set $d_1 = d_2 = .4$, then *p*-curve estimates $\hat{d} = .4$. But *p*-curve also estimates $\hat{d} = .4$ if we set $d_1 = .3$ and $d_2 = .5$, or $d_1 = .2$ and $d_2 = .6$, such that the average true effect is .4. *p*-curve is robust to heterogeneity in effect size across studies (see Supplement 2 for more details and additional variations).

9. Readers wanting even more details can see the R code behind Figure 3 in Supplement 4.

10. Note that we already showed *p*-curve to be robust to heterogeneity in effect size. We focus on homogenous studies not to ensure that *p*-curve is valid, but to ensure our benchmark for truth is.

11. Their I^2 statistic across labs, which measures the percent of variance explained by lab heterogeneity (Higgins, Thompson, Deeks, & Altman, 2003), was <10% for sunk costs, and <.01% for Asian disease. Neither of the effects were associated with significant differences between American and non-American labs, nor between laboratory and online methods of data collection (all *p*s > .29; see Table 3 in Klein et al., 2014).

12. The sunk cost problem consisted of a question asking participants to rate their willingness to attend a match of their favorite team on a freezing cold day either if they had paid for a ticket or if it was free (for earlier studies of this effect, obtaining bigger effects probably due to more precise wording, see Arkes & Blumer, 1985; Thaler, 1980). The Asian disease problem consisted of a binary hypothetical choice problem that presented participants with the same information using either a gains or a losses frame (Tversky & Kahneman, 1981).

13. Two of the 36 labs had disproportionately large sample sizes, $N > 1{,}000$. To make things more interesting, the results reported in the main text exclude those two labs so that there is more variability (we use 34 different labs). Results with those two labs are reported in the next footnote.

14. Figure 5 was constructed excluding the two largest labs. With them included, the effect size estimates for the sunk cost fallacy are $\hat{d}_{\text{naive}} = .33$, $\hat{d}_{\text{Earnest}} = .28$, $\hat{d}_{p\text{-curve}} = .28$. For the Asian Disease these are $\hat{d}_{\text{naive}} = .63$, $\hat{d}_{\text{Earnest}} = .60$, $\hat{d}_{p\text{-curve}} = .60$. We deemed these results less interesting because the very large samples greatly reduce bias in the subset of $p < .05$ studies.

15. The average test statistic is $\frac{2.12+2.43}{2} = 2.28$, leading to $t(38) = 2.28$, $p = .028$.

16. The reason not to include average *p* values nor average test statistics is that they are not uniform under the null. For instance, as one averages more and more *p* values the result converges to .5 (rather than to a uniform distribution of 0 to 1).

17. It is easy to specify *p*-curve in a way that allows for effects of opposite sign to be included simultaneously, but we decided against it. The reason is that it is quite unlikely that a true underlying effect leads to significant results of opposite sign. The probability of getting a $p < .05$ effect of the "wrong" sign is necessarily smaller than 2.5% (that's the probability if $d = 0$). If a study is powered to just 20%, the odds are less than 1 in 1,000, and if a study is powered to 50%, the odds are less than 1 in 11,000. If statistically significant opposite sign effects are observed, separate analyses for those $\hat{d} > 0$ and $\hat{d} < 0$ are almost surely more meaningful and informative. Our choice-overload demonstration exemplifies how splitting effects by sign can lead to insightful inferences.

18. In the Appendix, we explain the procedure and provide R Code for a difference of means *t* test. Because the mapping between a test results (e.g., *t* value) and effect size (e.g., Cohen's *d*) is different for each design, users of p-curve will need to either create custom programs for different designs (e.g., interactions, regressions, mixed designs), or more conveniently, use p-curve to estimate average power and then convert the average power to a measure of effect size. Supplement 4 includes the R program that computes average power.

19. Recall that we assume all *t* values are of the same sign, so we do not count significant effects of the wrong sign into power calculations.

20. Differences across *p*-curves for the same level of power depend on the degrees of freedom of the *t* test (on how thick the tails of the *t* distribution are), and converge to the normal distribution's *p*-curve as the degrees of freedom increase.

References

Arkes, H. R., & Blumer, C. (1985). The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, *35*, 124–140.

Ashenfelter, O., Harmon, C., & Oosterbeek, H. (1999). A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour Economics*, *6*, 453–470.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.

Chapman, G., & Johnson, E. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes*, *79*, 115–153.

Chernev, A., Bockenholt, U., & Goodman, J. (2010). Commentary on Scheibehenne, Greifeneder, and Todd choice overload: Is there anything to it? *Journal of Consumer Research*, *37*, 426–428. doi:10.1086/655200

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cole, L. C. (1957). Biological clock in the unicorn. *Science*, *125*, 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*, 286–300.

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89–98.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*, 328–331.

Gerber, A. S., & Malhotra, N. (2008). Publication bias in empirical sociological research do arbitrary significance levels distort published results? *Sociological Methods & Research*, *37*, 3–30.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics*, *9*, 61–85.

Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 145–174). Chichester, England: Wiley.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557.

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*, *53*, 11–22.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–646.

John, L., Loewenstein, G. F., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524–532.

Klein, R. A., Ratliff, K., Vianello, M., Reginald, B., Adams, J., Bahnik, S., . . . Nosek, B. A. (2014). *Investigating variation in replicability: A "Many Labs" Replication Project*. Retrieved from Open Science Framework osf.io/wx7ck

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*, 107–112.

Oppenheimer, D., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical psychology*, *58*, 646–656.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis*. West Sussex, England: Wiley.

Scheibehenne, B., Greifendeder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research*, *37*, 409–425.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

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.

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., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.

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*, *54*, 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*, *49*, 108–112.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *1*, 39–60.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Wallis, W. A. (1942). Compounding probabilities from independent significance tests. *Econometrica*, *10*, 229–248.