Detecting Evidential Value and *p*-Hacking With the *p*-Curve Tool

A Word of Caution

Edgar Erdfelder¹ and Daniel W. Heck^{1,2}

¹Department of Psychology, School of Social Sciences, University of Mannheim, Germany

²Department of Psychology, University of Marburg, Germany

Abstract: Simonsohn, Nelson, and Simmons (2014a) proposed *p*-curve – the distribution of statistically significant *p*-values for a set of studies – as a tool to assess the evidential value of these studies. They argued that, whereas right-skewed *p*-curves indicate true underlying effects, left-skewed *p*-curves indicate selective reporting of significant results when there is no true effect (“*p*-hacking”). We first review previous research showing that, in contrast to the first claim, null effects may produce right-skewed *p*-curves under some conditions. We then question the second claim by showing that not only selective reporting but also selective nonreporting of significant results due to a significant outcome of a more popular alternative test of the same hypothesis may produce left-skewed *p*-curves, even if all studies reflect true effects. Hence, just as right-skewed *p*-curves do not necessarily imply evidential value, left-skewed *p*-curves do not necessarily imply *p*-hacking and absence of true effects in the studies involved.

Keywords: publication bias, *p*-hacking, false-positive results, *p*-curve, ANCOVA

One of the major problems in behavioral science is publication bias, defined by van Aert, Wicherts, and van Assen (2016, p. 713) as “the tendency for studies with statistically significant results to be published at a higher rate than studies with results that are not statistically significant” (see also, e.g., Greve, Bröder, & Erdfelder, 2013; Kühberger, Fritz, & Scherndl, 2014; Rosenthal, 1979; Rothstein, Sutton, & Borenstein, 2005; Ulrich, Miller, & Erdfelder, 2018). Notably, this problem is not unique to psychology, but applies to all empirical sciences that derive substantive conclusions from statistical tests (e.g., Banks, Kepes, & McDaniel, 2012; Ioannidis, 2005, 2008; Sterling, 1959). As a consequence, published research in any of these fields may represent a distorted picture of reality, with considerably more false-positive findings than expected under the $\alpha = .05$ type-1 error rate typically assumed in statistical inference (e.g., Button et al., 2013; Erdfelder & Ulrich, 2018; Ferguson & Heene, 2012; Fiedler, 2011; Francis, 2012, 2014; Pashler & Harris, 2012; Schimmack, 2012; Simmons, Nelson, & Simonsohn, 2011; Ulrich et al., 2016; Vul, Harris, Winkielman, & Pashler, 2009). Thus, to enhance the validity of meta-analytic reviews, it is crucial to find convincing answers to the question how to discriminate between sets of significant results that are likely to indicate real effects (i.e., true positives) and those that

merely reflect selective reporting of significant results when there is in fact no real effect (i.e., false positives).

Simonsohn, Nelson, and Simmons (2014a) proposed *p*-curve as a new tool to address this problem. As indicated by the 321 citations in the Web of Science as of April 23, 2019, Simonsohn et al.’s (2014a) method has become very popular in the few years since its publication. The *p*-curve method considers exact tail probabilities under the null hypothesis of published significant test results only – the so-called *p*-values. It essentially involves plotting a histogram and evaluating the distribution of independent *p*-values observed in a predefined set of studies that meet a number of inclusion criteria, for example, studies that (1) tested a substantive hypothesis of interest (e.g., power posing improves performance; see Carney, Cuddy, & Yap, 2015), (2) resulted in test statistics significant at $\alpha = .05$ (hence, $0 < p < .05$), and (3) were published in certain journals within a fixed period (see Cuddy, Schultz, & Fosse, 2018; Simmons & Simonsohn, 2017, for applications to the power-posing example). If all actually conducted tests of the hypothesis of interest (or a random sample thereof) were also reported, then the distribution of significant *p*-values observed in these studies reveals whether they contain or lack evidential value and thus corroborate or refute, respectively, the hypothesis of interest (Hung,

O'Neill, Bauer, & Köhne, 1997). Under the null hypothesis of no effect in the respective statistical model (in our example $H_0: \mu_1 = \mu_2$, with means referring to independent and identically normally distributed performance scores in the power-posing vs. control condition, respectively), all *p*-values for the observed *t* (or *F*) statistics are equally likely by definition. This also holds for significant *p*-values only, that is, conditional on $p < .05$. By implication, the true *p*-curve must be a flat line, and the observed frequency distribution of *p*-values must not differ significantly from a uniform distribution. Under the alternative hypothesis of an effect in the predicted direction ($H_1: \mu_1 > \mu_2$), in contrast, very small *p*-values close to zero will be more likely than larger *p*-values close to .05. Hence, the *p*-curve will be right-skewed under H_1 , and the degree of skewness monotonically increases with statistical power (Hung et al., 1997; see Simonsohn, Nelson, & Simmons, 2014b, p. 668, Figure 1, for an illustration, and Ulrich & Miller, 2018, for a precise formal description of *p*-curves).

Whereas both flat and right-skewed *p*-curves can occur given unbiased reporting of all test results (depending on whether H_0 or H_1 holds, respectively), left-skewed *p*-curves cannot occur under these scenarios. So how can left-skewed *p*-curves emerge? Simonsohn et al. (2014a) showed both formally and by means of Monte Carlo simulations that most forms of so-called *p*-hacking – repeatedly trying different ways to analyze the data until a significant result is obtained – not only inflates the factual type-1 error rate beyond the nominal α level (e.g., Simmons et al., 2011) but may also produce left-skewed *p*-curves in many cases. The precise form of *p*-hacking – whether by stopping data collection upon obtaining a significant result and continuing otherwise (“data peeking”), by trying different dependent variables and selecting the one that produces a significant result (“multiple testing”), by excluding outliers until a significant result is achieved (“data trimming”), or by repeatedly trying different statistical models or tests until a significant outcome is observed – does not matter much (see Ulrich & Miller, 2015; Ulrich et al., 2016, for overviews on variants of *p*-hacking). Any of these questionable research practices – or any combination thereof (see Simmons et al., 2011) – may produce left-skewed *p*-curves.

Using data peeking after every set of five participants per group in a *t*-test scenario as an example, Simonsohn et al. (2014a, p. 537, Figure 1) showed by means of simulations that the resulting *p*-curve is in fact markedly left-skewed when H_0 holds and that it still shows some degree of left-skewness even when H_1 holds with low to medium effect sizes. Only when the true effect size exceeds Cohen's $d = .50$, the *p*-curve resulting from data peeking becomes right-skewed, in line with what we would expect for nonselective reporting of true effects. Bishop and Thompson

(2016) presented similar simulation results and also showed that the correlation between subsequent *p*-values is crucial for obtaining left-skewness as a consequence of *p*-hacking under H_0 (see also Simonsohn et al., 2014a, Supplementary Material 3; Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2016; Lakens, 2015).

Simonsohn and collaborators suggested relying on statistical tests (i.e., Fisher's and Stouffer's method) to assess the discrepancy of the distribution of observed *p*-values from a specific null hypothesis (e.g., a uniform distribution under H_0). Based on the test outcomes, a *p*-curve can then be classified as showing (1) right-skewness, (2) flatness (i.e., flatter than expected given power = .33), or (3) left-skewness (Simonsohn et al., 2014a, p. 537; Simonsohn, Simmons, & Nelson, 2015, p. 1149). According to Simonsohn et al. (2014a, pp. 537–540), a significantly right-skewed *p*-curve indicates evidential value of the set of studies, a significantly flat *p*-curve indicates lack of evidential value, and a significantly left-skewed *p*-curve indicates both lack of evidential value and intense *p*-hacking.

We agree with Simonsohn and collaborators that the shape of the *p*-curve for a set of studies is an important piece of information that has largely been overlooked in meta-analytic research until recently. We also believe that the statistical machinery they developed for the analysis and evaluation of *p*-curves and implemented in the *p*-curve online app (see www.p-curve.com/app4) is helpful for practitioners. Nevertheless, despite this positive assessment, we are concerned about one key problem in the *p*-curve methodology and the simplistic interpretations suggested by Simonsohn et al. (2014a, 2015): the failure to distinguish between necessary and sufficient conditions for different shapes of *p*-curves. As shown by Simonsohn et al. (2014a) and summarized here, there is no doubt that unselected reporting of studies is a *sufficient* condition for right-skewness and flatness of *p*-curves, respectively, depending on whether a true effect exists or not. Likewise, there is no doubt that *p*-hacking is *sufficient* for left-skewness of *p*-curves under some conditions, as illustrated with the data peeking example above.

However, to allow the reverse conclusions suggested by Simonsohn et al. (2014a), namely, that right-skewness implies evidential value and that left-skewness implies both intense *p*-hacking and lack of evidential value, it is mandatory to prove that true effects and *p*-hacking are also *necessary* conditions for right-skewness and left-skewness of *p*-curves, respectively. Such a proof is missing. In other words, it remains to be investigated whether – apart from application errors, fraud, and honest errors (see Simonsohn et al., 2015) – right-skewed *p*-curves *can only* result from true effects whereas left-skewed *p*-curves *can only* result from *p*-hacking in the absence of true effects. In the following, we show that neither of these two statements is in

general logically valid and discuss the consequences of our results for (1) future *p*-curve analyses and interpretations, (2) the choice between different statistical tests of the same hypothesis in primary studies, and (3) reporting standards for statistical analyses.

Does a Right-Skewed *p*-Curve Imply Evidential Value?

Let us start with the question: Does a right-skewed *p*-curve imply the absence of *p*-hacking? Notably, this question has partly been answered already by Simonsohn et al. (2014a, p. 537, Figure 1G and H). In their data peeking simulation, they showed that *p*-hacking may underlie right-skewed *p*-curves, but only when the effect of interest is sufficiently strong (or *p*-hacking is mild enough) to “hide” the left-skewness that would be observed in case of a null effect. Thus, right-skewness does not imply absence of *p*-hacking when the set of studies contains strong evidential value for the hypothesis of interest. As Lakens (2015, p. 830, Figure 2) showed by means of simulations, this still holds when only 33% of the studies included in the *p*-curve contain strong evidential value. In other words, given that at least one third of the studies show strong evidential value, a markedly right-skewed *p*-curve results, even when *p*-hacked null effects underlie the remaining two thirds of the studies.

Now let us address a more fundamental problem, namely, whether right-skewness of *p*-curves can also result from *p*-hacking in the total absence of evidential value, that is, even if there is no true underlying effect in any of the studies involved. Somewhat surprisingly, this is indeed the case (Bruns & Ioannidis, 2016; Ulrich & Miller, 2015, 2018). For example, Ulrich and Miller (2015) point out that *p*-hacking cannot only result from sequential data-analysis strategies – performing several analyses one by one and stopping upon the first significant outcome – but also from parallel strategies – performing several analyses at once and then selecting the to-be-reported result that supports the hypothesis of interest most strongly. Given the accessibility of easy-to-use software packages that allow multiple analyses of the same data in a couple of minutes, the parallel-strategy assumption is plausible and in fact consistent with what some researchers have reported about their data analysis practices (for a summary, see Ulrich & Miller, 2015, pp. 1138–1139).

Assuming parallel data analysis strategies, there are many variants of *p*-hacking that may produce right-skewed *p*-curves in the absence of any effect. For example, selection of the variable from a set of dependent variables that produces the smallest *p*-value induces right-skewness,

irrespective of whether the dependent variables are correlated or not (Ulrich & Miller, 2015, pp. 1139–1140). For this strategy, the degree of skewness increases with the number *k* of dependent variables explored. Moreover, post-hoc construction of composite scores from a set of dependent variables may have the same effect (Ulrich & Miller, 2015, p. 1141).

Are there other strategies of selective reporting that may produce right-skewness in the absence of evidential value? Again, the answer is “yes.” Ulrich and Miller (2015, pp. 1141–1142) showed that, ironically, researchers inadvertently generate right-skewed *p*-curves if they honestly attempt to avoid α -error inflation by reporting an experiment (and all of its significant effects) only if at least two of *k* dependent variables produce significant effects (provided that these dependent variables are sufficiently correlated). The same holds when researchers are even more cautious and report an experiment only if all *k* analyzed dependent variables are significant in the predicted direction. Although these strategies cannot count as *p*-hacking (because they select entire experiments and not to-be-reported *p*-values), both strategies bias the *p*-curve toward an overrepresentation of very small *p*-values, giving rise to right-skewed curves even in the total absence of evidential value.

As recently shown by Ulrich and Miller (2018), right-skewed *p*-curves also emerge conditional on null effects when the likelihood to publish empirical work gradually increases for decreasing *p*-values below .05. In fact, based on a survey experiment among more than a thousand members of the German Psychological Society, Ulrich and Miller (2018, Table 2) found that researchers are more inclined to submit empirical work for publication when it results in $p = .0032$ (in this case 52% of the researchers are willing to submit) rather than $p = .0232$ (48.6% submissions) or $p = .0432$ (only 33.8% submissions), although $p < .05$ holds for all three scenarios.

Still another reason why right-skewed *p*-curves may result from null effects is omitted variable bias in observational research. Bruns and Ioannidis (2016) showed by means of simulations that *t*-tests of regression coefficients can produce markedly right-skewed *p*-curves under H_0 if an additional predictor of the dependent variable is ignored in the model. Notably, given a sufficiently large sample size, this also holds when the effect of the ignored predictor is so small that it most likely remains undetected with significance tests.

In sum, without additional assumptions or precautions, a right-skewed *p*-curve in itself neither implies the absence of *p*-hacking nor does it imply evidential value in any of the studies that is included in the *p*-curve. A *p*-curve can be significantly right-skewed even if it is based on null effects exclusively.

Does a Left-Skewed *p*-Curve Imply *p*-Hacking?

As we have seen in the previous section, right-skewness of *p*-curves is ambiguous. Is left-skewness of *p*-curves less ambiguous? To address this question, let us first investigate Simonsohn et al.'s (2014a, p. 539) demonstration example for left-skewed *p*-curves in more detail. Using a number of predefined selection rules, Simonsohn and collaborators chose 20 randomized experiments published in the *Journal of Personality and Social Psychology* for which results were reported only with an Analysis of Covariance (ANCOVA) based on a covariate observed prior to the experimental manipulation (see Simonsohn et al., 2014a, p. 539 and Supplementary Material 5). When hypothesizing that this set "was likely to have been *p*-hacked (and thus less likely to contain evidential value)," their implicit assumption was that many authors of these ANCOVA studies first tried a *t*-test or an Analysis of Variance (ANOVA) *F*-test and switched to the ANCOVA only when obtaining a nonsignificant outcome in the first step (see Simmons et al., 2011). In line with this idea, Simonsohn and collaborators indeed observed a monotonically increasing, significantly left-skewed *p*-curve.

Note, however, that selective reports of significant ANCOVAs can emerge in different ways, not just by sequential *p*-hacking. One option is the reversed sequential strategy: Starting with an ANCOVA and switching to an ANOVA (or *t*-test) if the former is significant, just to see whether the experimental effect can also be established with a more popular statistical test that is easier to communicate and requires less assumptions. If the ANOVA also becomes significant, then only this latter test will be reported for simplicity, resulting in a biased ANCOVA *p*-curve from which significant ANOVAs have been excluded. Still an alternative possibility is that researchers routinely perform ANCOVAs and ANOVAs in parallel. Both strategies – the reversed sequential strategy and the parallel strategy – are particularly likely in case of randomized experiments for which covariates exist that are known to affect the dependent variable of interest. Because of randomization, experimental conditions, and covariates must be stochastically independent. Hence, mean comparisons with and without the covariate – ANCOVAs and ANOVAs – refer to exactly the same null hypothesis (e.g., $H_0: \mu_1 = \mu_2$ in the 2-groups case). Notably, although both procedures assess equivalent null hypotheses given randomization, they may of course differ in statistical power, the more so the stronger the effect of the covariate on the dependent variable.

Given these facts and given the high prevalence of multiple analyses of the same data (cf. Ulrich & Miller, 2015), it seems likely that researchers often run both ANOVAs and ANCOVAs, provided the relevant literature suggests a particular covariate. If both tests turn out to be significant, however, it seems unlikely that all researchers would report both results.¹ To avoid redundancy and to comply with the succinctness requirements of reviewers and editors, some authors will report a single test only. If we now focus on those studies where only ANCOVA results are reported – as Simonsohn et al. (2014a) did – it is important to keep in mind that a representative *p*-curve would require that all ANCOVA values with $p < .05$ are included, irrespective of the outcome of the ANOVA test. However, this is very unlikely to happen. If the ANCOVA and the ANOVA result in, say, $p = .011$ and $p = .039$, respectively, and the author chooses not to report both results, then she or he will almost surely report the ANOVA result only, most likely with the idea in mind that the result is significant anyway and one should therefore report the "standard test" for randomized experiments that requires less journal space and is easier to communicate. Although this choice is comprehensible and free from *p*-hacking, the consequence of reporting the ANOVA only is that the significant ANCOVA $p = .011$ is excluded from the ANCOVA *p*-curve. The same happens to other cases where both ANCOVA and ANOVA are significant.

Because ANCOVAs outperform ANOVAs in terms of statistical power when the effect of the covariate on the dependent variable is moderate or strong, ANCOVA *p*-values tend to be smaller than ANOVA *p*-values when there is a true effect, resulting in a selective dropout of very small *p*-values from the ANCOVA *p*-curve. Based on these considerations, it seems possible that left-skewed ANCOVA *p*-curves can emerge even if all *p*-values are based on true effects and *p*-hacking is not involved, simply by switching from ANCOVA-reports to ANOVA-reports by default whenever both procedures are significant. If so, the conclusion would be that left-skewness of *p*-curves does not necessarily imply *p*-hacking in the absence of a real effect.

Simulation 1: ANCOVA *p*-Curves When Reports of ANOVAs or *t*-Tests Are Preferred

We performed a Monte Carlo simulation study to show that left-skewed ANCOVA *p*-curves may result from excluding

¹ Note that even if they did, we would not see the corresponding *p*-values in the ANCOVA *p*-curve analysis of Simonsohn et al. (2014a), as this analysis excludes all cases where both ANCOVA and ANOVA results were reported.

significant *p*-values whenever the corresponding ANOVA *p*-value is also significant. Using the R programming environment (R Core Team, 2018), we generated samples consistent with the assumptions of the General Linear Model for a *t*-test scenario and $n = 20$ per group with versus without an additional covariate. Thus, errors were assumed to be independent and identically normally distributed (and generated with the *mvrnorm* function of the R package MASS; Venables & Ripley, 2002). True mean differences between groups and standard deviations were selected to define four types of mean differences, namely, a scenario consistent with H_0 (i.e., Cohen's $d = 0$) and three scenarios consistent with H_1 (specifically, Cohen's $d = .493$, $.636$, and $.909$). For a two-tailed *t*-test (or the equivalent ANOVA *F*-test) without a covariate these three effect sizes are associated with a statistical power of $1 - \beta = .33$, $.50$, and $.80$, respectively, given $\alpha = .05$ and $n = 20$ per group (calculated with G*Power 3.1.9.4; Faul, Erdfelder, Buchner, & Lang, 2009). For analyses with covariates, small, medium, and large correlations between covariates and dependent variables were considered (i.e., $r = .10$, $.30$, and $.50$, see Cohen, 1988). In line with the principle of randomization, experimental conditions and covariates were stochastically independent in all simulations. Cross-classification of four types of mean differences with the three covariate correlations generates $4 \times 3 = 12$ simulation scenarios in total. For each of these 12 scenarios, 50,000 Monte Carlo random

samples were generated and analyzed with both a standard two-tailed *t*-test that ignores the covariate and, in addition, an ANCOVA that includes the covariate (see Electronic Supplementary Material (ESM) 1 for the R script we used).

Results

Figure 1 illustrates the *p*-curves derived from the 50,000 Monte Carlo samples per scenario for two different test strategies, namely, (1) ANCOVA *p*-curves with all significant *p*-values included, irrespective of the outcome of the *t*-test, and (2) ANCOVA *p*-curves observed when all cases with a significant outcome of the *t*-test are excluded. Obviously, the unselected *p*-curves for ANCOVAs (first row of Figure 1) behave as they should according to the *p*-curve logic: They are flat when there is no real effect, and they are right-skewed when there is an effect, the more so the larger the power to detect the effect. We also see that *p*-curve skewness increases slightly with the covariate-criterion correlation, indicating the increase in power that is gained by including the covariate in the linear model. In fact, looking at Monte Carlo estimates of the ANCOVA power for this strategy (see Table 1), we see that ANCOVA power is larger than the power of the two-tailed *t*-test, provided that the correlation between covariate and dependent variable exceeds $r = .10$.

Figure 1. Monte Carlo-simulated *p*-curves (50,000 simulations per curve) for a 2-groups ANCOVA given different effect sizes d with associated power values of $.33$, $.50$, and $.80$, and covariate-criterion correlations r with (a) all $p < .05$ included (first row) and (b) all $p < .05$ included except those with a significant outcome of the corresponding *t*-test (second row). Dotted black lines indicate perfectly flat *p*-curves.

Table 1. 2-groups ANCOVA Monte Carlo power estimates for different scenarios

Covariate correlation	Two-tailed <i>t</i> -test scenario			
	H_0 holds	$H_1: (1-\beta) = .33$	$H_1: (1-\beta) = .50$	$H_1: (1-\beta) = .80$
$r = 0.10$	0.051	0.326	0.494	0.794
$r = 0.30$	0.050	0.349	0.528	0.823
$r = 0.50$	0.049	0.413	0.606	0.888

Note. The covariate correlation is the true correlation between the covariate and the dependent variable. Experimental conditions and the covariate are stochastically independent. The different *t*-test scenarios refer to effect sizes of $d = 0$ (H_0 holds), $d = 0.493$ ($H_1: (1-\beta) = .33$), $d = 0.636$ ($H_1: (1-\beta) = .50$), and $d = 0.909$ ($H_1: (1-\beta) = .80$).

The pattern looks entirely different for ANCOVA *p*-values that remain after exclusion of studies that only report the significant *p*-values of the preferred *t*-test (second row in Figure 1). In line with the prediction derived in the previous section, most of these *p*-curves are markedly left-skewed, despite the fact that *p*-hacking is not involved here. Note that this also holds when the curves are based on real effects in the underlying population, with the sole exception of flat *p*-curves for medium or large covariate correlations when the power of the *t*-test is high. Notably, even in the latter cases, the *p*-curves do not exhibit the right-skewed pattern one would expect given a true effect in the population. Rather, these curves are flat which would lead to the incorrect interpretation of a “lack of evidential value.”

Discussion

Simulation 1 showed that left-skewed ANCOVA *p*-curves can emerge in the absence of *p*-hacking and even if the underlying data contain evidential value, in direct contrast to the interpretation of Simonsohn et al. (2014a). The single plausible assumption required to derive this result is that researchers prefer reporting ANOVAs (or *t*-tests) over ANCOVAs whenever both tests are significant. Even in those cases where such a preference for reporting *t*-tests does not suffice to produce left-skewed curves (i.e., when both true effects and covariate-criterion correlations are medium or large) the guidelines for interpreting *p*-curves would still be misleading because they suggest “lack of evidential value” when in fact strong evidence underlies all *p*-values.

Notably, our results in the second row of Figure 1 differ markedly from apparently similar simulations Simonsohn et al. (2014a) published in the Supplementary Materials of their paper (Section 7, Figure S2). With their simulations, the authors tried to address a related concern raised by the first author of the current paper (EE) when serving as a reviewer of Simonsohn et al.’s (2014a) manuscript. However, it is important to note that the simulations

illustrated in their Figure S2 are based on the assumption that between 25% and 75% of the studies with significant ANCOVAs and ANOVAs are reported as ANCOVAs only (thus ignoring the significant *t*-test/ANOVA outcome). Our simulations, in contrast, assume that no researcher reports only the ANCOVA result when both tests are significant. We believe that our assumption is more reasonable and realistic, simply because reporting the standard *t*-test (or the corresponding ANOVA) is more customary, easier, and takes less journal space (since it involves less assumptions) than reporting only the significant ANCOVA or both test results jointly.

In sum, our simulations show that it is possible to obtain left-skewed *p*-curves by switching from ANCOVA to *t*-test reports whenever both tests turn out to be significant. Since such a behavior cannot reasonably be criticized as *p*-hacking (because *p*-hacking involves suppression of non-significant test statistics rather than suppression of significant outcomes), this example shows that *p*-hacking is not implied by left-skewness of *p*-curves. In other words, not only right-skewed but also left-skewed *p*-curves are ambiguous with respect to evidential value and *p*-hacking, respectively. Proving that *p*-hacking is involved in a set of studies requires more than just showing that the corresponding *p*-curve is left-skewed.

A possible objection against our argument is that ANCOVA *p*-curves are rather special and perhaps not representative for applications of the *p*-curve tool in general. Most *p*-curve applications aim at evaluating the evidential value for or against a substantive hypothesis and typically integrate *p*-values across different types of tests, not just ANCOVAs. We agree but maintain that this does not invalidate our argument. The danger of misinterpreting skewness of *p*-curves is a general one and not limited to ANCOVA *p*-curves. More precisely, whenever researchers perform two tests of the hypothesis of interest and choose to report only the less powerful test when either test is significant, *p*-curves tend to become flat or perhaps left-skewed, even if true effects underlie the data. We will investigate this in more detail in a second simulation.

Simulation 2: *p*-Curves for the *t*-Test When Reports of *U*-Tests Are Preferred

Consider 2-groups *t*-tests and Wilcoxon-Mann-Whitney *U*-tests as another possible example. Under the General Linear Model for 2-groups designs, both tests assess the same null hypothesis $H_0: \mu_1 = \mu_2$, with the power of the *t*-test exceeding the power of the *U*-test if H_0 is false (Faul

et al., 2009). If researchers prefer to report only the *U*-test and not the *t*-test when both are significant (e.g., because of the weaker distributional assumptions involved in *U*-tests) – a preference that is typical for many sciences such as medicine or the neurosciences – the *p*-curve for *t*-tests tends to become left-skewed, since small *p*-values for *t*-tests are excluded.

We performed a second Monte Carlo simulation to illustrate this effect. Using the R programming environment (R Core Team, 2018) we generated samples consistent with the General Linear Model for a *t*-test scenario and $n = 20$ per group, again using the *mvrnorm* function of the R package MASS (Venables & Ripley, 2002). Resembling Simulation 1, true mean differences between groups and standard deviations were selected to define four types of mean differences, that is, a scenario consistent with H_0 (i.e., Cohen's $d = 0$) and three scenarios consistent with H_1 (specifically, Cohen's $d = .493$, $.636$, and $.909$, resulting in power values of $1 - \beta = .33$, $.50$, and $.80$, respectively, given $\alpha = .05$ and $n = 20$ per group of the 2-groups *t*-test). For each of these four simulation scenarios, 50,000 Monte Carlo random samples were generated and analyzed with both a standard two-tailed *t*-test and the corresponding Wilcoxon-Mann-Whitney *U*-test (see ESM 2 for the R script we used).

Results

Figure 2 illustrates the *t*-test *p*-curves derived from the 50,000 Monte Carlo samples per scenario (1) with all cases $p < .05$ included (first row), and (2) significant *t*-tests when all cases with a significant outcome of the corresponding *U*-test are excluded (second row). Again, the unselected *p*-curves for *t*-tests in the first row behave as they should according to the *p*-curve logic. The pattern looks entirely different for *t*-test *p*-values that remain after exclusion of significant outcomes of the *U*-test (second row in Figure 2). All of these *p*-curves are markedly left-skewed, even when true effects are very strong.

Discussion

Conceptually replicating Simulation 1, our second simulation showed that left-skewed *p*-curves for the *t*-test can emerge in the absence of *p*-hacking and even if the underlying data are based on strong true effects, conflicting with the interpretation of left-skewed *p*-curves suggested by Simonsohn et al. (2014a). The single plausible assumption required in this case is that researchers prefer reporting

U-tests over reporting corresponding *t*-tests whenever both tests are significant.

General Discussion

Building on previous work by Ulrich and Miller (2015, 2018) and Bruns and Ioannidis (2016), the current paper shows that neither evidential value nor lack of evidential value or *p*-hacking can be inferred from the shape of the *p*-curve for a set of studies. The first and foremost implication of this insight is that conclusions about the evidential value of a set of studies should be much more cautious than has heretofore been standard. This holds for both *p*-curve analyses that apparently support (e.g., Lakens, 2017, professor priming; Steiger & Kühberger, 2018; Weintraub, Hall, Carbonella, Weisman De Mamani, & Hooley, 2017) or contradict (e.g., Lakens, 2017, elderly priming; Vadillo, Gold, & Osman, 2016; van Aert et al., 2016) a substantive hypothesis of interest. As previously shown by Bruns and Ioannidis (2016) and Ulrich and Miller (2015, 2018), right-skewed *p*-curves can result from selective reporting of test results when there is no true effect in any of the studies involved. As shown here, selective nonreporting unrelated to *p*-hacking may also produce left-skewed *p*-curves and – perhaps more frequently – flat *p*-curves, when true effects underlie all test results (see the lower parts of Figures 1 and 2 in particular). Taken together, these examples prove that, in general, right-skewness of *p*-curves does not imply true effects in the underlying studies, flatness does not imply null effects, and left-skewness does not imply *p*-hacking of null effects.

Of course, this somewhat sobering conclusion does not mean that *p*-curves are useless. The *p*-curve method can indeed show that some selection process of unknown origin² must underlie a set of significant *p*-values whenever a left-skewed *p*-curve is observed. Such a finding is remarkable and requires an explanation. Just as evidential value is one of several plausible explanations for right-skewed *p*-curves, *p*-hacking is only one of several plausible explanations for left-skewed ones. However, a plausible explanation is far from being a logically sound proof, which is why such explanations should be handled with caution (see van Aert et al., 2016, for a similar conclusion).

Is there any way to gain deeper insights into the unknown mechanisms and data analysis strategies underlying the shape of *p*-curves? The first idea that comes to mind is to simulate different possible mechanisms and strategies and to see which one produces a shape of the *p*-curve that is

² We cannot even be sure that the origin lies in the selective reporting of results or test statistics in the original research. It might well lie in the selection process of the meta-analyst who decides about inclusion and exclusion rules for *p*-curves. We thank Richard Morey for pointing this out.

Figure 2. Monte Carlo-simulated *p*-curves (50,000 simulations per curve) for a 2-groups *t*-test given different effect sizes *d* and associated power values of .33, .50, and .80, with (a) all $p < .05$ included (first row) and (b) all $p < .05$ included except those with a significant outcome of the corresponding *U*-test (second row). Dotted black lines indicate perfectly flat *p*-curves.

closest to the observed one. Unfortunately, this idea is of no help in general because different strategies may result in the same *p*-curve. For example, the *p*-curves in the second row of Figure 1 represent significant ANCOVAs with all cases of a significant *t*-tests for the same data excluded. This describes the selection principle underlying the data, but it does not imply anything about the mechanism that caused it. It could be sequential *p*-hacking as presupposed by Simonsohn et al. (2014a) (i.e., first performing a *t*-test and proceeding to the ANCOVA only if the *t*-test is non-significant), but it could also be a reversed sequential strategy, or ANCOVA and *t*-test analyses conducted in parallel, followed by preferred *t*-test reports when both tests are significant.

We believe the only way to corroborate an interpretation of the *p*-curve's shape is to go beyond the statistical evidence contained in the *p*-curve. For instance, if sequential *p*-hacking is the actual mechanism causing the bias in the ANCOVA example, the covariate is selected on an ad-hoc basis without any strong theoretical or empirical reasons to include it into the model a priori. The opposite is true for reversed sequential and parallel-analyses strategies. These strategies can only count as proper strategies if there are reasons to include the covariate into the model on a priori grounds. The content of the articles from which the *p*-values were derived will in general reveal which of these two possibilities is more likely. If the authors disclose

their reasons for inclusion of the covariate, and if these reasons are in line with (and supported by) the relevant literature, then this refutes the *p*-hacking interpretation. Note, however, that this is qualitative evidence, not quantitative evidence such as the test statistics proposed to assess the *p*-curve's shape. Without additional qualitative evidence, researchers should refrain from insisting on a particular *p*-curve interpretation that is not logically implied by a set of sufficient conditions.

A key problem appears to be that researchers typically assess the same hypothesis based on the same data in multiple ways, either sequentially or in parallel. Often, the idea behind multiple analyses of the same data is to establish an effect of interest as significant using a procedure that requires minimal assumptions and is easiest to defend and to communicate. For example, for randomized designs, a significant ANOVA or *t*-test is often preferred to a significant ANCOVA (cf. Simulation 1), and a significant Kruskal-Wallis *H*-test or Mann-Whitney *U*-test might even be more attractive than an ANOVA or *t*-test (cf. Simulation 2). Similarly, researchers may often prefer to report conservative tests such as Fisher's exact test or *p*-values corrected for multiple comparisons, among others. However, conservative tests usually also have a lower statistical power and thus result in larger *p*-values on average. Since the *p*-curve methodology focuses on the subset of *p*-values truncated to the interval between 0 and .05, *p*-curves cannot discriminate between

(a) *p*-values that are generally biased upwards due to a preference for more conservative tests and (b) *p*-values that are biased upwards toward the common publication threshold of $p < .05$ as a consequence of *p*-hacking.

Can we avoid selection biases in reported test statistics in the future or at least make them unlikely? Instead of selectively reporting only the most conservative test if *p*-values lead to identical conclusions across multiple tests, authors should be encouraged to focus on a single, a priori defined test of the hypothesis of interest – the optimal test. Thereby, selection biases are effectively prevented, provided that authors comply with this rule. For instance, preregistration of studies requires researchers to define the optimal statistical test a priori before any data are collected and thereby effectively prevents *p*-hacking (Greve et al., 2013; Nosek & Lakens, 2014).

However, what is the “optimal test” of the hypothesis of interest? If several tests of the same null hypothesis are available – as in case of ANCOVAs and ANOVAs for randomized designs – then the apriori most powerful test should be chosen if its assumptions hold for the given scenario (e.g., Erdfelder, 2010). Clearly, if there is a covariate that is known to exert a nontrivial effect on the dependent variable under scrutiny (i.e., $r > .10$) and that is unrelated to the manipulated factor, then ANCOVA is the method of choice. Thus, whenever these conditions are met, authors should refrain from ANOVAs for randomized designs and perform and report ANCOVAs only. In fact, the most recent edition of the statistical guidelines of the Psychonomic Society (2019) endorses the same conclusion in Point 2 C: If “(.) it is inappropriate to analyze data without a covariate, then re-analyze those same data with a covariate and report only the latter (..).” It is thus very unfortunate that ANCOVAs are discredited as potential *p*-hacking instruments when in fact they are the most valuable instruments available to test many hypotheses about differences in means. Just as ANOVAs are not always better than ANCOVAs for randomized designs, a general preference for nonparametric tests is unsubstantiated, even if these tests may appear less problematic than parametric tests because of their weaker distributional assumptions. Numerous simulation studies have shown that ANOVAs are quite robust both under H_0 and under H_1 (e.g., Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). Hence, if there is no gross violation of the assumption of independent and normally distributed errors and homogeneous variance, then *t*-tests and ANOVAs are known to be more powerful than Wilcoxon-Mann-Whitney *U*-tests or Kruskal-Wallis *H*-tests, respectively. In these cases, preference should be given to the former, not to the latter.

Of course, conducting and reporting supplemental analyses in addition to the focal test (e.g., in an Appendix) is permissible and may provide useful information.

Additional analyses are especially useful if researchers are interested in establishing the robustness of their conclusions against alternative ways of testing the hypothesis of interest or if exploratory use of the data is intended in addition to the assessment of a core hypothesis. However, we deem it to be important that the optimal test, that is, the arguably most powerful test of the hypothesis of interest given reasonable statistical assumptions, is defined a priori (ideally, in a preregistration) and reported by default in the main text. Suppression of such focal tests and replacement by less powerful alternatives might not only produce meta-analytic anomalies such as left-skewed *p*-curves as illustrated here. Perhaps even more importantly, this practice would further reduce the low level of statistical power that has plagued behavioral research since decades and contributed to the excessively high rate of false-positive results in the published literature (Ulrich et al., 2016).

Another important point concerns the reporting standards in psychological publications. The publication manual of the American Psychological Association requires that all tests and analyses conducted for a set of data should be reported, at least in Supplementary Online archives assigned to journal articles (American Psychological Association, 2010, p. 34). Many authors do not seem to take this seriously. For this reason, we recommend inclusion of a data analysis declaration in the publication process, similar to the ethics declarations that have been mandatory for quite some time. In this data analysis declaration, authors should certify that they reported all their data analyses at least in summary form, either in the paper or in Appendices and Supplementary Materials (hence, an extended version of the “21 word solution” proposed by Simmons, Nelson, & Simonsohn, 2012). The same point was made in the American Statistical Association’s statement on *p*-values: “Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted and all *p*-values computed” (Wasserstein & Lazar, 2016, p. 10). Having access to the results of all analyses conducted for a specific set of data would enable us to complement *p*-curves with results not reported in the paper. Hence, open and transparent science – often discussed as an effective measure to overcome the replication crisis in psychology (e.g., Glöckner, Fiedler, & Renkewitz, 2018; Schönbrodt, Maier, Heene, & Bühner, 2018) – is also an effective instrument to improve the empirical basis of *p*-curves, helping us to learn more about the factual evidential value provided by a set of studies.

As a final point, we like to point out that the ambiguities of the *p*-curve method we criticized here also affect its use

to adjust meta-analytic effect-size estimates for publication bias (Simonsohn et al., 2014b). Hence, critical assessments of this use are warranted (van Aert et al., 2016) and alternative meta-analytic methods to estimate the extent of publication bias and to correct effect-size estimates accordingly must not be ignored (Carter, Schönbrodt, Gervais, & Hilgard, 2019; McShane, Böckenholt, & Hansen, 2016; Ulrich et al., 2018).

To conclude, we have shown that any *p*-curve result – if considered in isolation – is ambiguous with respect to the evidential value and/or *p*-hacking in the set of studies involved. To remove this ambiguity and to test certain interpretations of the *p*-curve's shape, we recommend qualitative (i.e., content-based) analyses of the articles from which the *p*-values were drawn. With respect to data analysis strategies in primary studies, we argue against post-hoc selection of the to-be-reported statistical result(s) from multiple data analyses addressing the same statistical hypothesis. Instead, we recommend a focal-test strategy, that is, conducting and reporting an apriori defined (ideally preregistered) most powerful test of a statistical hypothesis of interest (while additional analyses may still be provided in an Appendix). Most importantly, however, authors must disclose their data analysis strategies and report all analyses they conducted along with the results obtained.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at <https://doi.org/10.1027/2151-2604/a000383>

ESM 1. R code for *p*-curve ANCOVA vs. *t*-test

ESM 2. R code for *p*-curve *t*-test vs. *U*-test

References

- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.
- Banks, G. C., Kepes, S., & McDaniel, M. A. (2012). Publication bias: A call for improved meta-analytic practice in the organizational sciences. *International Journal of Selection and Assessment*, 20, 182–197. <https://doi.org/10.1111/j.1468-2389.2012.00591.x>
- Bishop, D. V. M., & Thompson, P. A. (2016). Problems in using *p*-curve analysis and text-mining to detect rate of *p*-hacking and evidential value. *PeerJ*, 4, e1715. <https://doi.org/10.7717/peerj.1715>
- Bruns, S. B., & Ioannidis, J. P. A. (2016). *p*-Curve and *p*-hacking in observational research. *PLoS One*, 11, e0149144. <https://doi.org/10.1371/journal.pone.0149144>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14, 365–376. <https://doi.org/10.1038/nrn3475>
- Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2015). Review and summary of research on the embodied effects of expansive (vs. contractive) nonverbal displays. *Psychological Science*, 26, 657–663. <https://doi.org/10.1177/0956797614566855>
- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2, 115–144. <https://doi.org/10.1177/2515245919847196>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cuddy, A. J. C., Schultz, S. J., & Fosse, N. E. (2018). *P*-curving a more comprehensive body of evidence on postural feedback reveals clear evidential value for power-posing effects: Reply to Simmons and Simonsohn (2017). *Psychological Science*, 29, 656–666. <https://doi.org/10.1177/0956797617746749>
- Erdfelder, E. (2010). Experimental Psychology: A note on statistical analysis. *Experimental Psychology*, 57, 1–4. <https://doi.org/10.1027/1618-3169/a000001>
- Erdfelder, E., & Ulrich, R. (2018). Zur Methodologie von Replikationsstudien [On the methodology of replication studies]. *Psychologische Rundschau*, 69, 3–21. <https://doi.org/10.1026/0033-3042/a000387>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, 7, 555–561. <https://doi.org/10.1177/1745691612459059>
- Fiedler, K. (2011). Voodoo correlations are everywhere – not only in neuroscience. *Perspectives on Psychological Science*, 6, 163–171. <https://doi.org/10.1177/1745691611400237>
- Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, 19, 975–991. <https://doi.org/10.3758/s13423-012-0322-y>
- Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, 21, 1180–1187. <https://doi.org/10.3758/s13423-014-0601-x>
- Glöckner, A., Fiedler, S., & Renkewitz, F. (2018). Belastbare und effiziente Wissenschaft: Strategische Ausrichtung von Forschungsprozessen als Weg aus der Replikationskrise [Sound and efficient science: A strategic alignment of research processes as a way out of the replication crisis]. *Psychologische Rundschau*, 69, 22–36. <https://doi.org/10.1026/0033-3042/a000384>
- Greve, W., Bröder, A., & Erdfelder, E. (2013). Result-blind peer-reviews and editorial decisions: A missing pillar of scientific culture. *European Psychologist*, 18, 286–294. <https://doi.org/10.1027/1016-9040/a000144>
- Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of *p*-values smaller than .05 in psychology: What is going on? *PeerJ*, 4, e1935. <https://doi.org/10.7717/peerj.1935>
- Hung, J. H. M., O'Neill, R. T., Bauer, P., & Köhne, K. (1997). The behavior of the *p*-value when the alternative hypothesis is true. *Biometrics*, 53, 11–22. <https://doi.org/10.2307/2533093>
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, 686–701. <https://doi.org/10.1371/journal.pmed.0020124>
- Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19, 640–648. <https://doi.org/10.1097/EDE.0b013e31818131e7>
- Kühnberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between

- effect size and sample size. *PLoS One*, 9, e105825. <https://doi.org/10.1371/journal.pone.0105825>
- Lakens, D. (2015). What *p*-hacking really looks like: A comment on Masicampo and LaLonde (2012). *The Quarterly Journal of Experimental Psychology*, 68, 829–832. <https://doi.org/10.1080/17470218.2014.982664>
- Lakens, D. (2017). Professors are not elderly: Evaluating the evidential value of two social priming effects through *p*-curve analyses. *PsyArXiv Preprints*. <https://doi.org/10.31234/osf.io/3m5y9>
- McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11, 730–749. <https://doi.org/10.1177/1745691616662243>
- Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, 45, 137–141. <https://doi.org/10.1027/1864-9335/a000192>
- Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7, 531–536. <https://doi.org/10.1177/1745691612463401>
- Psychonomic Society. (2019). *Statistical guidelines*. Madison, WI. Retrieved from <https://www.psychonomic.org/page/statisticalguidelines>
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, 86, 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, UK: Wiley.
- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17, 551–566. <https://doi.org/10.1037/a0029487>
- Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology*, 6, 147–151. <https://doi.org/10.1027/1614-2241/a000016>
- Schönbrodt, F. D., Maier, M., Heene, M., & Bühner, M. (2018). Forschungstransparenz als hohes wissenschaftliches Gut stärken: Konkrete Ansatzmöglichkeiten für Psychologische Institute [Fostering research transparency as a key property of science: Concrete action options for psychology departments]. *Psychologische Rundschau*, 69, 37–44. <https://doi.org/10.1026/0033-3042/a000386>
- 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. <https://doi.org/10.1177/0956797611417632>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Dialogue*, 26, 4–7. <https://doi.org/10.2139/ssrn.2160588>
- Simmons, J. P., & Simonsohn, U. (2017). Power posing: *P*-curving the evidence. *Psychological Science*, 28, 687–693. <https://doi.org/10.1177/0956797616658563>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143, 534–547. <https://doi.org/10.1037/a0033242>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). *P*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9, 666–681. <https://doi.org/10.1177/1745691614553988>
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better *p*-curves: Making *p*-curve analysis more robust to errors, fraud, and ambitious *p*-hacking. A reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, 144, 1146–1152. <https://doi.org/10.1037/xge0000104>
- Steiger, A., & Kühnberger, A. (2018). A meta-analytic re-appraisal of the framing effect. *Zeitschrift für Psychologie*, 226, 45–50. <https://doi.org/10.1027/2151-2604/a000321>
- 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. <https://doi.org/10.1080/01621459.1959.10501497>
- Ulrich, R., Erdfelder, E., Deutsch, R., Strauß, B., Brüggemann, A., Hannover, B., ... Rief, W. (2016). Inflation von falsch-positiven Befunden in der psychologischen Forschung: Mögliche Ursachen und Gegenmaßnahmen [Inflation of false positive results in psychological research: Possible causes and countermeasures]. *Psychologische Rundschau*, 67, 163–174. <https://doi.org/10.1026/0033-3042/a000296>
- Ulrich, R., & Miller, J. (2015). *P*-hacking by post hoc selection with multiple opportunities: Detectability by skewness test? Comment on Simonsohn, Nelson and Simmons (2014). *Journal of Experimental Psychology: General*, 144, 1137–1145. <https://doi.org/10.1037/xge0000086>
- Ulrich, R., & Miller, J. (2018). Some properties of *p*-curves, with an application to gradual publication bias. *Psychological Methods*, 23, 546–560. <https://doi.org/10.1037/met0000125>
- Ulrich, R., Miller, J., & Erdfelder, E. (2018). Effect size estimation from *t* statistics in the presence of publication bias: A brief review of existing approaches with some extensions. *Zeitschrift für Psychologie*, 226, 56–80. <https://doi.org/10.1027/2151-2604/a000319>
- Vadillo, M. A., Gold, N., & Osman, M. (2016). The bitter truth about sugar and willpower: The limited evidential value of the glucose model of ego depletion. *Psychological Science*, 27, 1207–1214. <https://doi.org/10.1177/0956797616654911>
- van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on *p* values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science*, 11, 713–729. <https://doi.org/10.1177/174569161616650874>
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York, NY: Springer.
- Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzling high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4, 274–290. <https://doi.org/10.1111/j.1745-6924.2009.01125.x>
- Wasserstein, R. L. W., & Lazar, N. A. (2016). The ASA’s statement on *p*-values: Context, process, and purpose. *The American Statistician*, 70, 129–133. <https://doi.org/10.1080/00031305.2016.1154108>
- Weintraub, M. J., Hall, D. L., Carbonella, J. Y., Weisman De Mamani, A., & Hooley, J. M. (2017). Integrity of literature on expressed emotion and relapse in patients with schizophrenia verified by a *p*-curve analysis. *Family Process*, 56, 436–444. <https://doi.org/10.1111/famp.12208>

History

Received August 9, 2018

Revision received June 14, 2019

Accepted June 17, 2019

Published online December 20, 2019

Acknowledgments

The authors are grateful to Richard Morey for helpful comments on a previous version of this manuscript.

Funding

Manuscript preparation was supported by the Research Training Group “Statistical Modeling in Psychology,” funded by the Deutsche Forschungsgemeinschaft (GRK 2277).

Edgar Erdfelder

Cognition and Individual Differences Lab
University of Mannheim
Schloss, Ehrenhof-Ost
68131 Mannheim
Germany
erdfelder@uni-mannheim.de

<https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000383> - Wednesday, April 26, 2023 4:04:52 AM - IP Address: 54.210.77.194