

Conducting meta-analyses based on *p*-values: Reservations and recommendations for applying *p*-uniform and *p*-curve

Robbie C. M. van Aert¹, Jelte M. Wicherts¹, and Marcel A. L. M. van Assen^{1,2}

¹Tilburg University ²Utrecht University

Author Note

The preparation of this article was supported by Grants 406-13-050 and 452-11-004 from the Netherlands Organization for Scientific Research (NWO). We thank Hilde Augusteijn, Marjan Bakker, Paulette Flore, Gregory Francis, Chris Hartgerink, and Michèle Nuijten for their valuable comments on an earlier draft of this paper.

Correspondence concerning the article should be addressed to Robbie C. M. van Aert, Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, the Netherlands

Abstract

Because evidence of publication bias in psychology is overwhelming, it is important to develop techniques that correct meta-analytic estimates for publication bias. Van Assen, Van Aert, and Wicherts (2015) and Simonsohn, Nelson, and Simmons (2014a) developed *p*-uniform and *p*-curve, respectively. The methodology on which these methods are based has great promise for providing accurate meta-analytic estimates in the presence of publication bias. However, we show that in some situations *p*-curve behaves erratically while *p*-uniform may yield implausible negative effect size estimates. Moreover, we show that (and explain why) *p*-curve and *p*-uniform overestimate effect size under moderate to large heterogeneity, and may yield unpredictable bias when researchers employ *p*-hacking. We offer hands-on recommendations on applying and interpreting results of meta-analysis in general and *p*-uniform and *p*-curve in particular. Both methods as well as traditional methods are applied to a meta-analysis on the effect of weight on judgments of importance. We offer guidance for applying *p*-uniform or *p*-curve using R and a user-friendly web application for applying *p*-uniform (<https://rvanaert.shinyapps.io/p-uniform>).

Keywords: *p*-uniform, *p*-curve, meta-analysis, *p*-hacking, heterogeneity

Conducting meta-analyses based on p -values: Reservations and recommendations for applying p -uniform and p -curve

Meta-analysis is the standard technique to synthesize effect sizes of several studies on the same phenomenon. A well-known problem of meta-analysis is that effect size can be overestimated because of publication bias (e.g., Ioannidis, 2008; Lane & Dunlap, 1978). Publication bias is here defined as the tendency of studies with statistically significant results to be published at a higher rate than studies with results that are not statistically significant. Because evidence of publication bias is overwhelming across many scientific disciplines (Fanelli, 2012), it is important to develop techniques that correct the meta-analytic estimate for publication bias (Moreno, Sutton, Ades, et al., 2009). Recently, Van Assen, Van Aert, and Wicherts (2015) and Simonsohn, Nelson, and Simmons (2014a) have independently developed methods aiming to provide an accurate meta-analytic estimate in the presence of publication bias. Their methods, p -uniform and p -curve, respectively, both make use of the distribution of statistically significant results yet differ in implementation. The goals of this paper are to introduce and explain both methods and their differences, to provide straightforward recommendations for applying meta-analysis, and to formulate guidelines for applying and interpreting results of p -uniform and p -curve.

A Primer on P -uniform and P -curve

Simonsohn, Nelson, and Simmons (2014b) described how statistically significant p -values of studies on an effect could be used to test this effect against the null-hypothesis that the effect equals zero. This idea was not new; Fisher (1925, Chapter 4) already developed a method for testing the null-hypothesis of no effect by means of combining p -values. However, the novelty of p -curve lies with its use of only the statistically significant p -values, which are

arguably not affected by publication bias. The method was called *p*-curve because it analyzed the curve or distribution of *p*-values. The logic of *p*-curve suggests that there is no effect in the studies in the meta-analysis if the *p*-values are uniformly distributed (i.e., *p*-curve is flat), whereas there is an effect whenever the *p*-value distribution or *p*-curve is right-skewed (Hung, O'Neill, Bauer, & Köhne, 1997).

A disadvantage of *p*-curve at that time was that effect size could not be estimated. Van Assen et al. (2015) developed another method analyzing statistically significant *p*-values, called *p*-uniform, which is able to estimate the effect size in a set of studies. Van Assen et al. (2015) called their method *p*-uniform, because the effect size estimate is equal to the value for which the *p*-value distribution conditional on that value is uniform (as we explain below). Besides estimating the effect size, *p*-uniform can also estimate a confidence interval around the effect size estimate, includes a test of publication bias, and, similar to *p*-curve (Simonsohn, Nelson, & Simmons, 2014b), tests the null-hypothesis of no effect. Simonsohn et al. (2014a) later extended *p*-curve to also estimate effect size. However, *p*-curve neither provides a confidence interval nor a test for publication bias. In the present study, we will focus on effect size estimation by both *p*-curve and *p*-uniform.

Van Assen et al. (2015) and Simonsohn et al. (2014a) convincingly illustrated the strengths of *p*-uniform and *p*-curve and the logic upon which it is based for carrying out meta-analysis. They showed that the methods provide accurate effect size estimates in the presence of publication bias, even when the number of statistically significant studies is small. Similarly, both methods were found to perform well when studies have the same sample sizes, when studies differ in sample size, and in the scenario where there is (small) heterogeneity of effect size (i.e., when the underlying (population) effect sizes actually differ between studies in the meta-

analysis). Moreover, results of Simonsohn et al. (2014a) suggested that *p*-hacking, or the original researcher's use of strategies to achieve statistical significance (Simmons, Nelson, & Simonsohn, 2011), leads to an underestimation of effect size in analyses based on *p*-curve, whereas it leads to overestimation of effect size in traditional meta-analysis (Bakker, Van Dijk, & Wicherts, 2012).

Three Reservations

Although we are convinced of the potential and validity of the logic of *p*-uniform and *p*-curve, we add three important reservations to the application of the methods and the general methodology in its current state. More specifically, we first demonstrate that *p*-uniform and *p*-curve may yield implausible negative (*p*-uniform) or inaccurate (*p*-curve) estimates in meta-analyses with *p*-values close to the significance level (considered equal to .05 in the present paper). Second, we explain why and show that *p*-hacking does not always cause *p*-curve's and *p*-uniform's effect sizes to be underestimated as was stated in Simonsohn et al. (2014a). Finally, we show that, in contrast to the results in Simonsohn et al. (2014a), *p*-uniform and *p*-curve cannot deal with a substantial amount of heterogeneity (i.e., there is no single true effect size underlying the studies in the meta-analysis, but rather a distribution of true effect sizes). Based on our explanation of the methods and the reservations, we formulate recommendations for applying meta-analysis in general and interpreting results of *p*-uniform and *p*-curve in particular. These hands-on recommendations are summarized in Table 1. Scientists who consider using these methods have to be aware of conditions in which the methods should not be interpreted, or interpreted with caution.

Table 1. Recommendations for meta-analysis and applying *p*-uniform and *p*-curve

Recommendations for meta-analysis and applying *p*-uniform and *p*-curve:

1) Check for evidence of *p*-hacking in the primary studies

→ In case of strong evidence or strong indications of *p*-hacking, be reluctant with interpreting estimates of traditional meta-analytic techniques and *p*-uniform and *p*-curve, because their effect size estimates may be biased in any direction depending on the type of *p*-hacking.

2) Apply fixed-effect and random-effects meta-analysis, as well as *p*-uniform or *p*-curve, and report their results conforming to the Meta-Analysis Reporting Standards (MARS [American Psychological Association, 2012, Appendix])

3) Check for direct or indirect evidence of publication bias

→ In case of evidence of publication bias, interpret results of *p*-uniform or *p*-curve rather than of fixed-effect and random-effects meta-analysis; in the absence of such evidence, interpret results of fixed-effect and random-effects meta-analysis

4) Set the effect size estimate of *p*-uniform or *p*-curve equal to zero if the average *p*-value of the statistically significant studies is larger than .025

5a.) If effect size is homogenous or heterogeneity small to moderate ($I^2 < 0.5$), interpret the estimate of *p*-uniform and *p*-curve as estimates of the average *population* effect size; otherwise they overestimate average *population* effect size and should be interpreted as estimates of the average true effect size of only the set of statistically significant studies

5b) In case of substantial heterogeneity and if desired, create homogeneous subgroups of primary studies based on theoretical or methodological considerations in order to estimate with *p*-uniform and *p*-curve the average *population* effect size underlying the studies in each subgroup

In the remainder of the paper, we illustrate major issues involved in applying *p*-curve and *p*-uniform by considering a recent meta-analysis of studies on the effect of weight on judgment of importance (Rabelo, Keller, Pilati, & Wicherts, 2015). We will briefly describe other meta-analysis methods using statistically significant effect sizes, introduce the basic idea underlying *p*-uniform and *p*-curve, and illustrate the logic of and computations in of *p*-uniform and *p*-curve in Appendix A. The analyses that form the basis of our three reservations and recommendations are presented in the next sections. Readers who do not want to delve into the (technical) details of *p*-uniform and *p*-curve can skip these sections and move over to the Discussion and Conclusion section, where we explain the recommendations in Table 1. R code for all our analyses is available in the supplementary materials.

Example

Rabelo et al. (2015) conducted a meta-analysis on the effect of weight on judgments of importance. The theory underlying the studies included in the meta-analysis is that the physical experience of weight (e.g., holding a heavy object) influences how much importance people assign to things, issues, and people (IJzerman, Padiotis, & Koole, 2013; Jostmann, Lakens, & Schubert, 2009). For instance, in their second study, Jostmann et al. (2009) found that participants who held a heavy clipboard attributed more importance to fairness in decision-making as opposed to participants holding a light clipboard. Table B1 in the appendices provides the full references, sample sizes (n_i^1 and n_i^2), *t*-values, and *p*-values from the 25 studies of this kind published in the embodiment literature.

According to the first recommendation, we should consider the presence of *p*-hacking in the primary studies included in the meta-analysis. We believe that the studies on the link between weight and importance are mostly studies in which the specifics of the analysis are often neither

preregistered nor clearly restricted by theory. Hence, according to Recommendation 1, we would use caution in interpreting the current results and await new (preferably pre-registered) studies in this field.

Four different meta-analytic estimates of the (mean) effect size underlying the weight-importance studies are presented in Table 2. In line with Recommendation 2, we first fitted traditional fixed-effect and random-effects meta-analysis. Both analyses yielded the same effect size estimate of 0.571 (95% confidence interval: [0.468;0.673]), which is highly statistically significant ($z = 10.90, p < .001$) and suggests a medium to large effect of the experience of weight on how much importance people assign to things (see Table 2). *P*-uniform's publication bias test suggested that there is evidence for publication bias ($z = 5.058, p < .001$), so we should interpret the results of *p*-uniform or *p*-curve rather than the standard meta-analytic estimates (Recommendation 3). Because the average *p*-value of the 23 statistically significant studies equals .0281, we set the effect size estimate of *p*-uniform and *p*-curve equal to 0, in line with Recommendation 4. When not setting the estimate to 0, applying *p*-curve and *p*-uniform yields a nonsignificant *negative* effect size (see Table 2), and *p*-uniform's 95% confidence interval (-0.676; 0.160) suggests that the effect size is small at best.

Table 2. Results of *p*-uniform, *p*-curve, fixed-effect (FE) meta-analysis, and random-effects (RE) meta-analysis when applied to the meta-analysis reported in Rabelo et al. (2015) of the effect of weight on the judgment of importance in the moral domain.

	<i>p</i> -uniform	<i>p</i> -curve	FE meta-analysis	RE meta-analysis
Effect size estimate	-0.179	-0.172	0.571	0.571
95% CI	(-0.676;0.160)	-	(0.468;0.673)	(0.468;0.673)
Test of H ₀ : $\delta = 0$	$z=0.959; p=.831$	$\chi^2(46)=55.833; p=.848$	$z=10.904; p<.001$	$z=10.904; p<.001$
Pub. bias test	$z=5.058; p<.001$	-		-

The null-hypothesis of no heterogeneity among the included studies was not rejected ($Q(24) = 4.55, p = 1, I^2=0$), which suggests that *p*-uniform and *p*-curve may accurately estimate the average population effect size (Recommendation 5a). Note that due to the absence of heterogeneity, effect size estimates of fixed-effect and random-effects meta-analysis were identical. Although the lack of heterogeneity suggests that the effects are homogeneous, in this particular instance, homogeneity is excessive (with a *p*-value of the *Q*-test very close to 1). Such excessive homogeneity is unlikely to occur under normal sampling conditions (Ioannidis, Trikalinos, & Zintzaras, 2006) and could be caused by publication bias (Augusteijn, 2015), possibly in combination with *p*-hacking. Our preliminary conclusion about the effect of physical experience of weight on importance would be that there is as yet no evidence in the literature for such an effect.

Other Methods Using *P*-values for Estimation

Several other methods were developed that use *p*-values in order to obtain an effect size estimate corrected for publication bias. Hedges (1984) developed a method for correcting meta-analytic effect sizes for publication bias that is similar to *p*-uniform and *p*-curve. He derived the maximum likelihood estimator of effect size under a model with only statistically significant results and studied the bias in the effect size estimate. Although Hedges (1984) discussed the application to meta-analyses, he only examined the bias in effect size of one statistically significant study. Hedges' method and its performance is not further examined in this paper because it is currently not applied in practice.

Other methods for obtaining effect size estimates corrected for publication bias are selection models (Hedges & Vevea, 2005). Selection models use an effect size model and a weight function for correcting the effect size estimates for publication bias. The effect size model describes the distribution of effect sizes in case all studies get published. The weight function yields probabilities of observing a particular study given its effect size or *p*-value. Studies' effect sizes are then weighted by these probabilities in order to get an effect size corrected for publication bias (for an overview on selection models see Hedges & Vevea, 2005). Drawbacks of selection models are that they require a large number of studies (i.e., more than 100) in order to avoid non-convergence (e.g., Field & Gillett, 2010; Hedges & Vevea, 2005), often yield implausible weight functions (Hedges & Vevea, 2005), are hard to implement, and require sophisticated assumptions and difficult choices (Borenstein, Hedges, Higgins, & Rothstein, 2009, p. 281). A recently proposed alternative for selection models based on Bayesian statistics showed promising results and does not suffer from convergence problems when the number of studies in the meta-analysis is small (Guan & Vandekerckhove, 2015). However, a disadvantage of the latter method is that it makes stronger assumptions on weight functions than *p*-uniform

and *p*-curve. *P*-uniform and *p*-curve assume that the probability of publishing a finding is independent of its *p*-value given its statistical significance, whereas the models in the method described in Guan and Vandekerckhove (2015) assume specific weights of findings depending on their *p*-value, significant or not. Because both significant and nonsignificant *p*-values are included, this Bayesian method makes assumptions about the extent of publication bias, and its estimates are affected by the extent of publication bias. For these reasons, we also no longer discuss selection models and their properties.

Basic Idea Underlying *P*-uniform and *P*-curve

P-uniform and *p*-curve use the distribution of only the statistically significant *p*-values for estimating effect size, for at least two reasons. First, collecting unpublished studies without the existence of study (or trial) registers is often hard, and these unpublished studies may provide biased information on effect size just like published studies do (Ferguson & Brannick, 2012). Second, evidence for publication bias is overwhelming. For instance, researchers have estimated that at least 90% of the published literature within psychology contains statistically significant results (e.g., Bakker et al., 2012; Fanelli, 2012; Sterling, Rosenbaum, & Weinkam, 1995), yielding overestimated effect sizes (e.g., Ioannidis, 2008; Lane & Dunlap, 1978). Because most published findings are statistically significant, only a relatively small number of published but statistically nonsignificant studies (on average up to 10%) need to be omitted from meta-analyses by *p*-curve and *p*-uniform.

Both *p*-uniform and *p*-curve are founded on the statistical principle that the distribution of *p*-values conditional on the true effect size is uniform.¹ This same statistical principle underlies standard null-hypothesis significance testing, where the *p*-values are uniformly distributed when the true effect size equals zero. In contrast to null-hypothesis significance testing, *p*-values from

p -uniform and p -curve are computed conditional not only on an effect size of zero (which would yield a simple transformation of the traditional p -values), but also conditional on other effect sizes (in which case the conditional p -value is not a simple transformation of the traditional p -value anymore). The effect size estimate of p -uniform and p -curve represents the effect size for which the conditional p -values are uniformly distributed.² So what both procedures do is to find an underlying effect, compute for each study the (conditional) p -value given this effect, and subsequently check whether these conditional p -values show a flat (i.e., uniform) distribution, like they should if indeed the studies reflect that underlying effect. The assumptions of p -uniform and p -curve are that all statistically significant studies have the same probability of getting published and being included in the meta-analysis, and are statistically independent (i.e., they should not be based on the same sample) (Van Assen et al., 2015). We describe the logic underlying p -uniform and p -curve as well as how the conditional p -value and p -uniform's and p -curve's effect size estimate are computed in Appendix A.

P -curve and P -uniform Overestimate Effect Size if Heterogeneity is Moderate to Large

Simonsohn et al. (2014a) stated that p -curve provides accurate effect size estimates in the presence of heterogeneity, i.e., in cases where true effects underlying the studies' observed effects differ. In a blog post Simonsohn (2015) qualified this statement as follows; "if we apply p -curve to a set of studies it tells us what effect we expect to get if we run those studies again". In other words, applying p -curve (and p -uniform) to a set of studies yields an accurate estimate of the average true effect size of *this exact set of studies*. We note, however, that it may be impossible to run exactly the same studies again since there will always be differences in, for instance, the participants included in the studies and the context in which the studies were

conducted.

Because of the importance of its implications for the interpretation of *p*-curve's estimate, we provide a simple example with heterogeneous effect sizes. Assume that the true effect size is equal to either 0 or 1 and that both underlying effects are equally likely, implying an average true effect size $\mu = .5$. Also assume that both true effect sizes are investigated with the same number of studies with a huge sample size, implying 5% and 100% of studies with true effects equal to 0 and 1 are statistically significant, respectively. Because studies' sample sizes are huge, the observed effect sizes of statistically significant studies are equal to (a number very close to) 0 and 1. As a result, *p*-curve's estimate equals $(0.05 \times 0 + 1 \times 1) / 1.05 = .952$, which is indeed equal to the average underlying true effect size of all the statistically significant studies. However, it is *much* larger than the true population average of .5. Moreover, traditional random-effects meta-analysis will provide a more accurate estimate of true average effect size (i.e., less positively biased) than *p*-curve, even under extreme publication bias.

It is often unrealistic to assume homogeneous true effect sizes underlying primary studies in psychological meta-analyses (e.g., Borenstein et al., 2009, Chapter 13). Moreover, researchers often want to estimate the true effect size in the population instead of the average true effect size in the studies included in the meta-analysis. That is, meta-analysts wish to obtain an estimate of .5, rather than .952 in our example. The reason why *p*-curve overestimates effect size under heterogeneity is that studies with an underlying true effect of 0 have a lower probability to be statistically significant, such that these studies are underrepresented in the meta-analysis. In our example, studies with large true effect size are 20 times more likely to be included in the meta-analysis than those with a zero effect size. Finally, we note that in this simple example, we may deal with the heterogeneity rather easily if true effect size (0 or 1) is perfectly linked to an

observed dichotomous study characteristic; applying *p*-curve or *p*-uniform to studies of both groups (a so called subgroup analysis [e.g., Borenstein et al., 2009, Chapter 19]) yields the correct estimates of 0 and 1. We therefore recommend applying these methods to subgroups of studies based on the different levels of a moderator in order to create more homogeneous sets of studies (Recommendation 5b, Table 1). However, in other realistic situations, the causes of heterogeneity are not simply observed, and subgroup analysis will not completely solve the heterogeneity problem.

To illustrate the effect of heterogeneity of effect sizes on the (over)estimation of effect size by *p*-curve and *p*-uniform, we also ran a simulation study where we varied heterogeneity from moderate to large under the usual scenario where heterogeneity is modeled continuously using a normal distribution of true effects, which is commonly assumed in meta-analysis (Raudenbush, 2009). As in Simonsohn et al. (2014a), 5,000 studies with statistically significant results were generated on which the meta-analysis was conducted. All studies had two conditions with 50 cases each, with population variance equal to 1 in both conditions. Average population effect size was .397, and standard deviations of true effect size (denoted by τ) were 0, 0.2, 0.4, 0.6, and 1, roughly corresponding to I^2 (i.e., ratio of heterogeneity to total variance [Higgins and Thompson, 2002]) values of 0, .5 (moderate heterogeneity), .8 (large heterogeneity), .9, and .96 in the population of studies. Table 3 provides the estimates of *p*-curve, *p*-uniform, fixed-effect meta-analysis, and random-effects meta-analysis (with restricted maximum likelihood estimator for estimating the amount of heterogeneity) of all studies with a statistically significant positive effect. For *p*-uniform we used the Irwin-Hall estimator and the so-called “1-*p*” estimator, a variant based on Fisher’s method, because this estimator is least affected by extreme effect sizes, and therefore provides better estimates in case of heterogeneity (Van Assen et al., 2015).

Table 3: Estimates of effect size using *p*-curve, *p*-uniform with Irwin-Hall estimator (IH), *p*-uniform with “1-*p*” estimator, fixed-effect (FE) meta-analysis, and random-effects (RE) meta-analysis (using restricted maximum likelihood for estimating the amount of heterogeneity) under different levels of heterogeneity (true effect .397), based on 5,000 studies with statistically significant positive effects.

	$\tau = 0, I^2 = 0$	$\tau = .2, I^2 = .5$	$\tau = .4, I^2 = .8$	$\tau = .6, I^2 = .9$	$\tau = 1, I^2 = .96$
<i>p</i> -curve	.393	.530	.703	.856	1.094
<i>p</i> -uniform (IH)	.383	.535	.724	.874	1.110
<i>p</i> -uniform (“1- <i>p</i> ”)	.387	.522	.679	.776	.903
FE meta-analysis	.553	.616	.738	.875	1.104
RE meta-analysis	.553	.616	.743	.897	1.185

The first column confirms that *p*-curve and *p*-uniform provide accurate estimates under homogeneity (effect size estimates are close to the true effect size .397), whereas fixed-effect and random-effects meta-analysis (both .553) overestimate effect size. The other columns, however, show that both *p*-curve and *p*-uniform *overestimate* the mean population effect size of .397 for moderate to large heterogeneity, and that this bias increases with larger heterogeneity. Note that the bias of fixed-effect and random-effects meta-analysis also increases with larger heterogeneity, and exceeds the bias of *p*-curve and *p*-uniform in these cases. Although *p*-uniform’s “1-*p*” estimator provides the best estimates, its bias is still so large that we do not recommend applying the methodology in its current state to estimate the average population effect size in situations where moderate or large heterogeneity is present or suspected

(Recommendation 5a, Table 1).

For illustrative purposes, we show how *p*-curve and *p*-uniform may still diagnose heterogeneity by applying *p*-uniform to one simulated meta-analysis of 20 studies with the aforementioned specifications; mean population effect size equal to .397, and large heterogeneity ($\tau = 1$; $I^2 = .96$). *P*-uniform's "1-p" estimator yielded an effect size estimate of $\hat{\delta} = .795$. However, a comparison of the expected conditional *p*-values to the observed conditional *p*-values for $\hat{\delta} = .795$ in the probability or P-P plot in Figure 1 clearly indicated systematic misfit. Specifically, observed conditional *p*-values should be uniformly distributed, as the expected conditional *p*-values. That is, all dots should fall on or close to the diagonal. But, assuming a fixed effect size of .795, the observed conditional *p*-values were either (much) too small (dots below the diagonal to the left) or (much) too large (dots above the diagonal to the right), signifying a large effect size variance. In other words, deviations from the diagonal in the P-P plot may be used to diagnose heterogeneity of effect size.

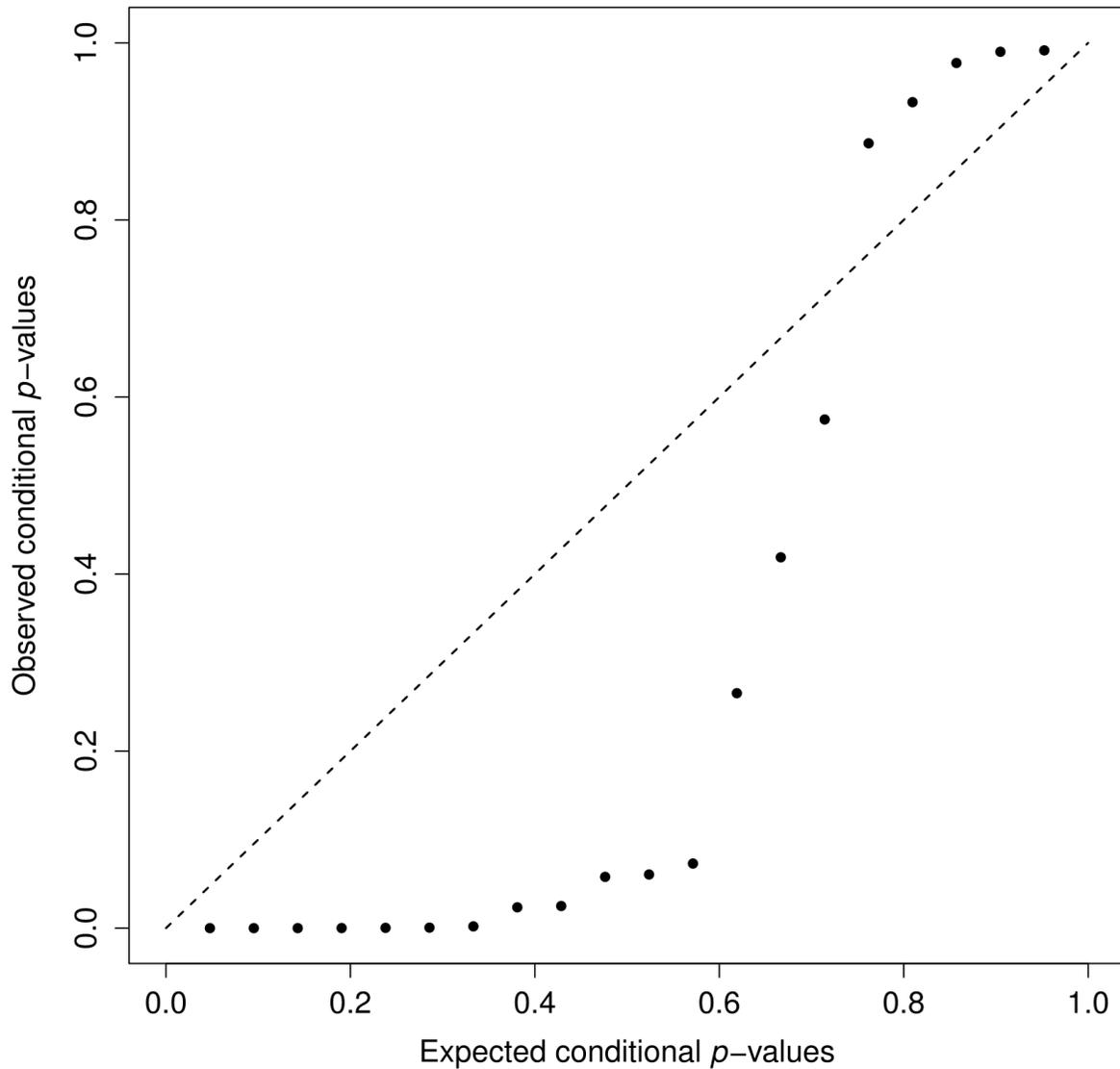


Figure 1. Probability or P-P plot for a meta-analysis of 20 studies with large heterogeneity.

To conclude, if moderate to large heterogeneity is present, then p -curve and p -uniform will estimate the average true effect underlying all *significant* studies in the meta-analysis. When the main goal of the meta-analysis is to estimate the average true effect of the *whole* population of studies in the presence of heterogeneity ($I^2 \geq .5$), we do not recommend using p -curve or p -uniform, because they then generally overestimate average true effect size (Recommendation 5a,

Table 1). As opposed to mainstream meta-analytic thinking, Simonsohn (2014a) argued that “the” average true effect size under heterogeneity often does not exist, and even that it is meaningless since studies cannot be run randomly. However, we believe the average true effect size may be meaningfully interpreted in the presence of heterogeneity in some situations, and consider heterogeneity to be both realistic for psychological studies (e.g., in 50% of the replicated psychological studies in the “Many Labs Replication Project,” heterogeneity was present [Klein et al., 2014]) and important to take into consideration when estimating average effect size.

Sensitivity to *P*-values Close to .05

Statistically significant *p*-values that are uniformly distributed in the interval (0; .05) are in line with a zero true effect size. Interestingly, a distribution of *p*-values with many *p*-values close to .05 (and say, an average *p*-value above .025) are not in line with a zero true effect size, but may indicate a *negative* true effect size. We will now show that if the majority of studies in the meta-analysis have a *p*-value just below the significance criterion of .05, then *p*-uniform yields implausible highly negative effect size estimates and a very wide confidence interval. Similarly, under these conditions *p*-curve will behave erratically.

To illustrate the consequences of many *p*-values just below .05 on the estimates of *p*-uniform and *p*-curve, consider doing a meta-analysis on the following three observed effect sizes with two conditions having equal sample sizes; Effect 1 with $d = .963$, $t(18) = 2.154$, $p = .045$ (two-tailed), Effect 2 with $d = .582$, $t(48) = 2.058$, $p = .045$, and Effect 3 with $d = .4$, $t(98) = 2.002$, $p = .048$. Several explanations exist for observing multiple *p*-values that barely pass the significance criterion as in this example. First, *p*-hacking such as optional stopping or data peeking (Hartgerink, Van Aert, Nuijten, Wicherts, & Van Assen, 2015; Lakens, 2014) or the

deletion of outliers to achieve statistical significance may yield a preponderance of *p*-values just below .05 (Bakker & Wicherts, 2014b). Another explanation is (bad) luck when the meta-analysis consists of a small number of studies and multiple studies coincidentally have *p*-values close to .05. The fixed-effect meta-analytic estimate for these three observed effect sizes is .506 ($p < .001$), with a 95% confidence interval excluding zero (.199, .812).³

Applying *p*-curve to this set of studies yields an effect size estimate of $d = -1.898$. Figure 2 displays the behavior of the Kolmogorov-Smirnov test statistic in *p*-curve with dots as a function of effect size. It shows that the Kolmogorov-Smirnov statistic in *p*-curve does not behave as it should (decrease to one minimum, and then increase, and being continuous for all effect sizes). This erratic behavior is caused by *p*-curve's implementation using the *t*-distribution from the software R (R Core Team, 2015), because R yields inaccurate probabilities for very high *t*-values in combination with an extreme non-centrality parameter (Witkovský, 2013). This inaccuracy may cause conditional *p*-values to be negative or undefined (division by zero), which yield the discontinuities in Figure 2. Therefore, *p*-curve's estimate cannot be trusted for this example.

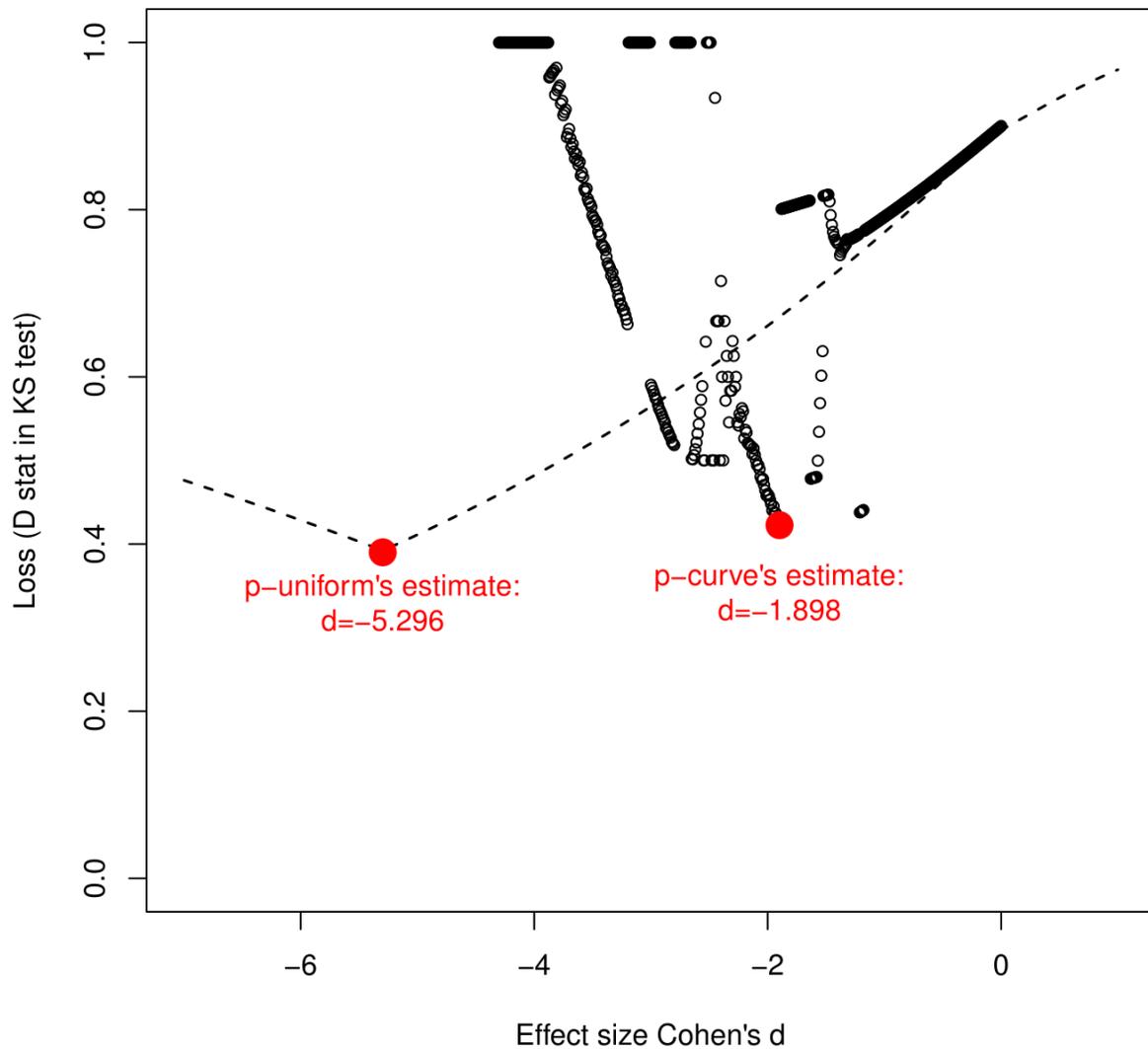


Figure 2. Values for Kolmogorov-Smirnov's test statistics in *p*-curve's and *p*-uniform's implementation for the example with three observed effect sizes and *p*-values close to .05.

P-uniform differs in implementation from *p*-curve because it uses the normal distribution instead of the *t*-distribution for computing conditional *p*-values. The studies' effect sizes are transformed into standardized effect sizes (Hedges' *g*) before the effect size is estimated.

Consequently, extreme tail probabilities can be computed, and therefore *p*-uniform behaves as it should, as can be seen from the dashed line in Figure 2. At the same time, *P*-uniform's estimate, also based on the Kolmogorov-Smirnov statistic to ease comparison with *p*-curve, is -5.296, which is clearly peculiar. Because a confidence-interval cannot be computed with the Kolmogorov-Smirnov statistic, we also calculated the Irwin-Hall estimates with *p*-uniform; $\hat{\delta} = -5.484$, 95% confidence interval (-15.219, -1.009). Although the behavior of *p*-uniform's estimator is correct, its effect size estimate (< -5) is unrealistically low; the probability of obtaining three *positive* statistically significant studies when $\delta = -5.484$ is essentially zero. Furthermore, *p*-uniform's confidence interval is very wide. We explain in the supplementary materials why these implausible negative estimates can be obtained and what can be concluded from these estimates. In order to deal with *p*-uniform's implausibly negative estimates and *p*-curve's erratic behavior, we recommend setting the effect size estimate of *p*-uniform and *p*-curve to zero in meta-analyses where the mean of the significant *p*-values of the primary studies is larger than .025 (Recommendation 4, Table 1). The cutoff of .025 is natural for two reasons. First, if the average *p*-value equals .025, *p*-uniform actually estimates $\hat{\delta} = 0$. Second, higher average *p*-values than .025 would yield negative effect size estimates, and testing is then redundant because the *p*-value of the test will be above .5 and hence cannot be statistically significant. The true effect size can, of course, be below zero, but a left-tailed hypothesis test is then required to examine whether the effect is smaller than zero.

***P*-hacking may cause bias in effect size estimates for *p*-uniform and *p*-curve**

Simonsohn et al. (2014a) examined the effect of *p*-hacking on *p*-curve's effect size estimation, considering three different *p*-hacking strategies; data-peeking, selectively reporting by using three dependent variables, and selectively excluding outliers. In data-peeking (or

optional stopping), observations are added whenever a test is not yet statistically significant. Their p -hacking strategy with multiple dependent variables refers to a practice where dependent variables are considered one by one, until one is found for which the test was statistically significant, which is then published. Selectively excluding outliers refers to deleting outliers whenever a test is not yet statistically significant. From their simulations of specific examples of these three practices, they concluded that p -curve *underestimates* effect sizes. However, p -hacking comprises a very large number of behaviors, and Simonsohn et al. (2014a) examined only three of these behaviors. We will now show that other types of p -hacking will lead to *overestimation* of effect size by p -curve and p -uniform.

As Simonsohn et al. (2014a, p.670) explain, p -hacking affects p -curve's estimate through the conditional p -value distribution. For instance, data peeking and selectively excluding outliers lead to a distribution with relatively more conditional p -values corresponding to just statistically significant results, which pulls p -curve's (and p -uniform's) estimate downward, as we have explained in the foregoing section. On the other hand, p -hacking behaviors yielding relatively more small p -values will result in an overestimation of effect size. Ulrich and Miller (2015) and Bruns and Ioannidis (2016) illustrate that multiple p -hacking behaviors may result in relatively more small p -values, which will lead to overestimation of effect size by p -curve (and p -uniform).

We examined the effect of two types of p -hacking on effect size estimation by p -curve and p -uniform. The first behavior again involves selectively reporting among three dependent variables, but differs from the procedure in Simonsohn et al. (2014a) in one crucial aspect; rather than reporting the *first* significant p -value, the *smallest* of three significant p -values is reported. The second behavior involves a "multiple conditions" scenario, where multiple experimental conditions are run and compared to the same control condition, and only the comparison yielding

the largest difference (and smallest *p*-value) is reported. We note that a large portion of surveyed psychologists have admitted to using at least once selective reporting among different dependent variables (63.4%) and not reporting all experimental conditions (27.7%) in their work (John, Loewenstein, & Prelec, 2012).

Figure 3 presents the estimates of *p*-uniform, as well as the true effect size and the effect size of fixed-effect meta-analysis (see the supplementary materials for the details of our simulations). We do not show *p*-curve's results because these are almost indistinguishable from *p*-uniform's results. Condition "First significant DV" and "Data peeking" are a replication of the simulations in Simonsohn et al. (2014a), showing that *p*-uniform and *p*-curve indeed underestimate effect size under these conditions. The estimate is slightly below the true effect size for "First significant DV", and about .2 lower on the scale of Cohen's *d* for "Data peeking" for all true effect sizes from 0 (no effect) to .8 (considered a large effect). Conversely, and as anticipated, both "DV with lowest *p*-value" and "Multiple conditions" overestimate effect size, and this overestimation increases for larger true effect sizes. What should also be mentioned is that *p*-uniform and *p*-curve did not always outperform traditional fixed-effect meta-analysis in the *p*-hacking scenarios we simulated. For instance, fixed-effect meta-analysis outperformed *p*-uniform and *p*-curve (i.e., presented less biased estimates) in the case of "Data peeking" (e.g., Francis, 2012; Van Aert, Maassen, Wicherts, & Van Assen, 2016). We therefore conclude that (i) *p*-hacking may bias *p*-uniform's and *p*-curve's estimate in any direction depending on the type of *p*-hacking, (ii) *p*-uniform's and *p*-curve's estimate are not *necessarily* better than those of fixed-effect meta-analysis when *p*-hacking occurs. Thus, *p*-uniform and *p*-curve can deal with publication bias, but (just like traditional fixed-effect and random-effects meta-analysis) neither corrects for *p*-hacking nor reacts predictably to it.

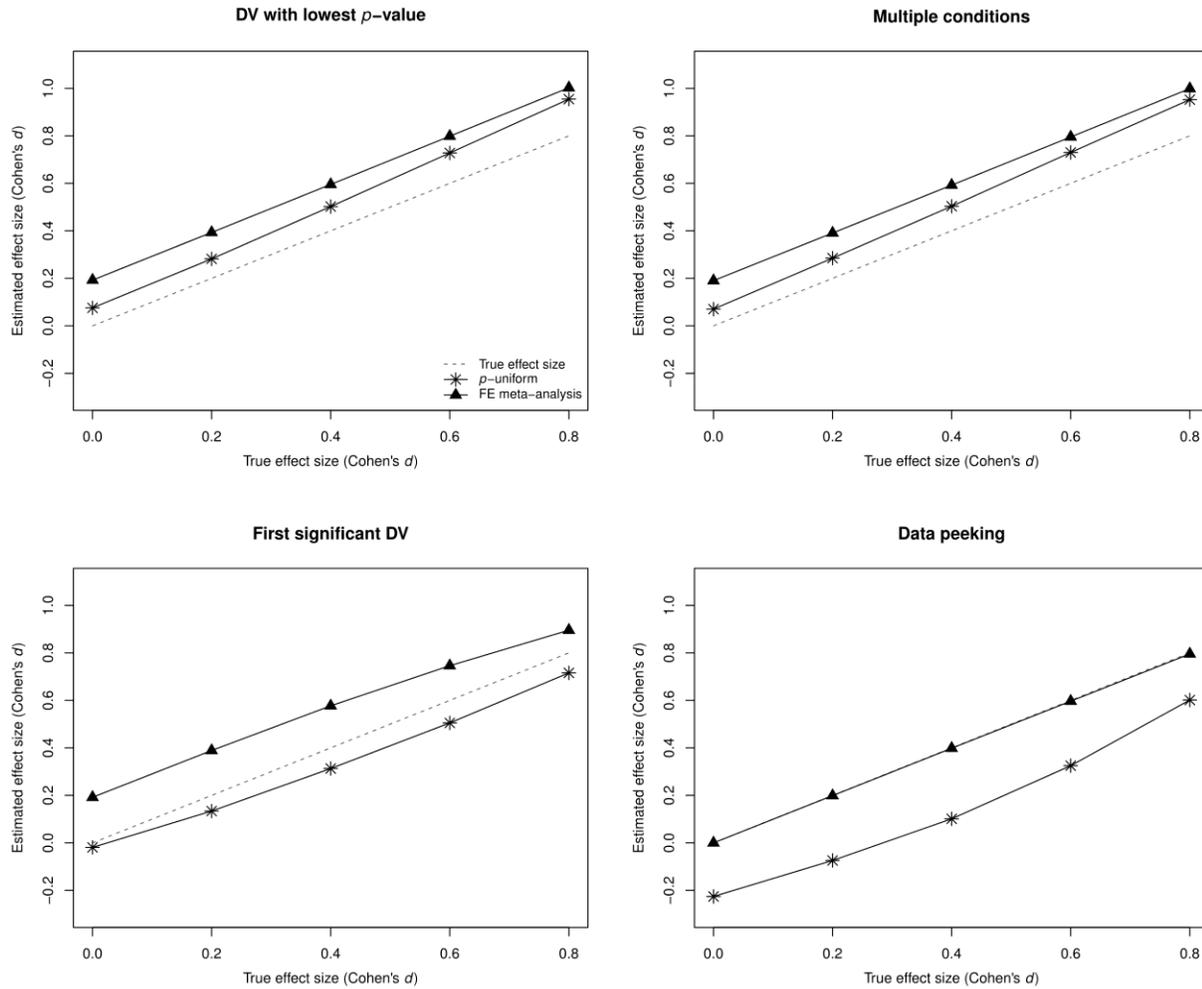


Figure 3. Effect size estimates of *p*-uniform and fixed-effect meta-analysis in case of four types of *p*-hacking.

Because the validity of results of both traditional meta-analysis methods and *p*-curve and *p*-uniform may be lowered by *p*-hacking, we recommend scrutinizing both data and studies included in the meta-analysis before applying meta-analytic methods. Underpowered primary studies (i.e., statistical power substantially below 0.8) and a preponderance of *p*-values just below .05 are signals for *p*-hacking. Other signals are unsystematic deletion of outliers and

reporting results of other than commonly used measurement instruments. If there are signs of *p*-hacking, we recommend applied researchers to be reluctant in interpreting the results of any meta-analysis (Recommendation 1, see Table 1).

Discussion and Conclusion

Recently, new methods were developed aiming to provide an accurate meta-analytic estimate in the presence of publication bias (Simonsohn et al., 2014a; Van Assen et al., 2015). These methods, *p*-uniform and *p*-curve, are based on the same basic idea but differ in implementation. The methods' idea is selecting only the statistically significant results and estimating the effect size using the principle of statistical theory that the distribution of (conditional) *p*-values based on the true effect size is uniform. Van Assen et al. (2015) and Simonsohn et al. (2014a) convincingly demonstrated the power of *p*-uniform and *p*-curve and the principles upon which the methods are based to carry out meta-analysis. In this paper, we explained the rationale and basics of both methods, added three reservations (concerning heterogeneity, incredible estimates, and *p*-hacking) to the application of both methods, and offered hands-on recommendations for researchers.

We explained that *p*-curve behaves erratically and yields inaccurate estimates in situations where multiple studies in a meta-analysis have *p*-values close to .05. Due to a different implementation, *p*-uniform does not suffer from this erratic behavior, but provides implausible negative estimates. These problems are solved by setting *p*-uniform's and *p*-curve's estimate to zero whenever the mean of statistically significant studies' *p*-values exceeds .025, i.e., whenever *p*-uniform's estimate is lower than zero. We also showed that *p*-hacking may bias *p*-uniform's and *p*-curve's estimate in any direction depending on the particular type of *p*-hacking, and these methods' estimates are not necessarily better than those of fixed-effect meta-analysis when *p*-

hacking has taken place. Finally, we explained that *p*-curve and *p*-uniform estimate the average true effect underlying all *significant* studies in the meta-analysis, but overestimate the average true effect of the *whole* population of studies whenever moderate to large heterogeneity is present.

On the basis of these and contemporary insights we formulated the recommendations summarized in Table 1. These recommendations hold for any meta-analysis and extend the Meta-Analysis Reporting Standards (MARS) as proposed by the APA (American Psychological Association, 2010, Appendix).

First, we recommend researchers to be reluctant to interpret the results of any meta-analytic technique if there are indicators for *p*-hacking in the primary studies (Recommendation 1), because *p*-hacking may bias the effect size estimates of meta-analysis in any direction. Indicators of potential *p*-hacking include the unsystematic deletion of outliers in many primary studies, the usage and reporting of multiple and different measures for the same dependent variable across primary studies, the common use of small underpowered studies, inconsistencies between sample size descriptions and degrees of freedom (Bakker & Wicherts, 2014a), and grossly misreported *p*-values (Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2015). *P*-hacking can be characteristic of a particular research field (e.g., different measures of dependent variables in a research field) as well as of a single study or a set of studies. Researchers can conduct a sensitivity analysis by comparing the results of traditional meta-analysis methods and *p*-uniform and *p*-curve with the results of these methods applied to only the studies where no *p*-hacking is suspected, for instance, because they involved the use of pre-registered data collection and analysis plans. Meta-analysts will probably observe indicators for *p*-hacking (if these are present) during the literature search and data extraction and do not have to go through all the

primary studies again to gather information about the potential presence of *p*-hacking.

Second, we recommend applying fixed-effect and random-effects meta-analysis and *p*-uniform or *p*-curve (Recommendation 2). The selection of a fixed-effect or random-effects meta-analysis should be based on whether a researcher wants to draw inferences on only the studies included in the meta-analysis (fixed-effect) or wants to generalize the meta-analytic results to the whole population of studies (random-effects) (see Borenstein, et al., 2009, Chapter 13 and Schmidt, Oh, & Hayes, 2009 for a more elaborate discussion on selecting fixed-effect or random-effects meta-analysis). Moreover, the estimate of fixed-effect meta-analysis, when compared to the estimate of random-effects meta-analysis, may signal publication bias; publication bias generally results in higher estimates of random-effects than fixed-effect meta-analysis because the studies with smaller sample sizes and usually overestimated effect sizes get less weight in fixed-effect meta-analysis (Greenhouse & Iyengar, 2009).

Next, we recommend checking for direct and indirect evidence of publication bias (Recommendation 3). Direct evidence can be obtained using the publication bias test in *p*-uniform. Previous research suggests *p*-uniform's test for publication bias has higher statistical power than traditional tests (Van Assen et al., 2015), which are known to have low statistical power (e.g., Borenstein et al., 2009, Chapter 30; Sterne & Egger, 2005). Moreover, use of the quite popular trim-and-fill method is discouraged because it often provides inaccurate results (Moreno, Sutton, Abrams, et al., 2009; Simonsohn et al., 2014a; Stanley & Doucouliagos, 2014; Van Assen et al., 2015). However, for a small number of studies in the meta-analysis or a small amount of publication bias, *p*-uniform's publication bias test will lack sufficient statistical power. In these cases, indirect evidence of publication bias may be used. An example of indirect evidence is if 80% or more of the primary studies' effect sizes are statistically significant when at

the same time these studies' sample sizes imply a power of .5 or less to detect a medium effect size (e.g., see Francis, 2013). In case of (direct or indirect) evidence of publication bias, we recommend that conclusions be based on *p*-uniform's or *p*-curve's results, rather than on fixed-effect and random-effects meta-analysis, because these traditional methods overestimate effect size in the presence of publication bias (e.g., Bakker, Van Dijk, & Wicherts, 2012; Ioannidis, 2008; Lane & Dunlap, 1978; Van Assen et al., 2015). Although *p*-uniform and *p*-curve also provide accurate effect size estimates even in the absence of publication bias (Simonsohn et al., 2014a; Van Assen et al., 2015), we recommend interpreting fixed-effect and random-effects meta-analysis in this case because these traditional methods yield more efficient and precise estimates.

We recommend setting *p*-uniform's and *p*-curve's estimate to 0 if the average *p*-value of statistically significant studies is larger than .025 (Recommendation 4); an average larger than .025 signals no evidence of an effect and/or the use of *p*-hacking in the set of included studies (in which case, meta-analytic methods' effect size estimation may be biased in any direction depending on the type of *p*-hacking; see Recommendation 1). Interpreting *p*-uniform's and *p*-curve's estimate as the average population effect size estimate is discouraged when effect size heterogeneity is large (Recommendation 5a). In this case, *p*-uniform's and *p*-curve's estimate reflects the average true effect underlying all *significant* studies in the meta-analysis. The average population effect size is overestimated (although the addition of *p*-hacking could complicate this pattern further) when there is moderate or large heterogeneity ($I^2 \geq .5$) and the average true effect of the *whole* population of studies is estimated. In order to deal with heterogeneous effect sizes and still be able to accurately estimate the average true effect of the *whole* population of studies, *p*-uniform or *p*-curve can be applied to homogeneous subgroups of

primary studies which were created based on theoretical (e.g., same population of participants being studied) or methodological considerations (using the same methodology, i.e. study design and measures) (Recommendation 5b). The implication of recommendations 3 and 5 is that, currently, no method provides accurate estimates of average population effect size in the presence of both publication bias and heterogeneity.

In the example meta-analysis described earlier, we applied *p*-uniform and *p*-curve to a set of primary studies on the effect of weight on judgment of importance (Rabelo et al., 2015). Researchers can also easily apply *p*-uniform or *p*-curve to their own data. User-friendly R code for applying *p*-uniform can be readily installed.⁴ Moreover, we developed a user-friendly web application for researchers who are not familiar with R (<https://rvanaert.shinyapps.io/p-uniform>). R code for estimating effect size with *p*-curve can be found in the supplementary materials of Simonsohn et al. (2014a). *P*-uniform has the advantage over *p*-curve that it also includes a publication bias test and yields a confidence interval around the effect size estimate.

To conclude, even though both *p*-uniform and *p*-curve are promising meta-analytic methods, the methodology underlying them is still under development, and properties of these methods still need to be examined under more stringent conditions (e.g., different forms of *p*-hacking). Moreover, both methods need to be extended to allow estimation of other effect sizes such as odds ratios, which have their own idiosyncrasies. Once the current methodology is further refined, particularly by enabling accurate estimation in case of heterogeneity, we believe it has the potential to become the standard meta-analytic tool correcting for publication bias. At present, however, researchers should follow the recommendations provided in Table 1 to avoid drawing erroneous conclusions from these still developing methods.

References

- Ackerman, J. M., Nocera, C. C., & Bargh, J. A. (2010). Incidental haptic sensations influence social judgments and decisions. *Science*, 328(5986), 1712-1715. doi:<http://dx.doi.org/10.1126/science.1189993>
- American Psychological Association. (2010). *Publication manual* (6th ed.). Washington, DC: American Psychological Association.
- Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268), 765-769.
- Augusteijn, H. E. M. (2015). *The effect of publication bias on the Q-test and assessment of heterogeneity* (Unpublished master's thesis). Tilburg University, Tilburg, the Netherlands.
- Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7(6), 543-554. doi:[10.1177/1745691612459060](https://doi.org/10.1177/1745691612459060)
- Bakker, M., & Wicherts, J. M. (2014a). Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS ONE*, 9(7): e103360. doi:[10.1371/journal.pone.0103360](https://doi.org/10.1371/journal.pone.0103360)
- Bakker, M., & Wicherts, J. M. (2014b). Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests: the power of alternatives and recommendations. *Psychological Methods*, 19(3), 409-427. doi:[10.1037/met0000014](https://doi.org/10.1037/met0000014)
- Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Belmont, CA: Duxbury.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons, Ltd.
- Bruns, S. B., & Ioannidis, J. P. (2016). P-curve and p-hacking in observational research. *PLoS ONE*, 11(2): e0149144. doi:[10.1371/journal.pone.0149144](https://doi.org/10.1371/journal.pone.0149144)
- Chandler, J. J., Reinhard, D., & Schwarz, N. (2012). To judge a book by its weight you need to know its content: Knowledge moderates the use of embodied cues. *Journal of Experimental Social Psychology*, 48(4), 948-953. doi:[10.1016/j.jesp.2012.03.003](https://doi.org/10.1016/j.jesp.2012.03.003)

- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891-904. doi:10.1007/s11192-011-0494-7
- Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120-128. doi:10.1037/a0024445
- Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 665-694. doi:10.1348/000711010X502733
- Fisher, R. A. (1925). *Statistical methods for research workers* (1st ed.). London: Oliver & Boyd.
- Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*(5), 153-169. doi: 10.1016/j.jmp.2013.02.003
- Greenhouse, J. B., & Iyengar, S. (2009). Sensitivity analysis and diagnostics. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 417-433). New York: Russell Sage Foundation.
- Guan, M., & Vandekerckhove, J. (2015). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin and Review*, Advance online publication. doi:10.3758/s13423-015-0868-6
- Gentle, J. E. (2004). *Random number generation and Monte Carlo methods* (2nd ed.). New York: Springer.
- Häfner, M. (2013). When body and mind are talking. Interoception moderates embodied cognition. *Experimental Psychology*, *60*(4), 255-259. doi:10.1027/1618-3169/a000194
- Hartgerink, C. H. J., Van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & Van Assen, M. A. L. M. (2015). Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*:e1935. doi: <https://doi.org/10.7717/peerj.1935>
- Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 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: Prevention, assessment, and adjustments*. Chichester: UK: Wiley.

- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539-1558. doi:10.1002/sim.1186
- Hung, 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*(1), 11-22.
- IJzerman, H., Padiotis, N., & Koole, S. L. (2013). Replicability of social-cognitive priming: The case of weight as an embodiment of importance. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2255726.
- Ioannidis, J. P. A., Trikalinos, T. A., & Zintzaras, E. (2006). Extreme between-study homogeneity in meta-analyses could offer useful insights. *Journal of Clinical Epidemiology*, *59*(10), 1023-1032. doi:10.1016/j.jclinepi.2006.02.013
- Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640-648. doi:10.1097/EDE.0b013e31818131e7
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524-532. doi:10.1177/0956797611430953
- Jostmann, N. B., Lakens, D., & Schubert, T. W. (2009). Weight as an embodiment of importance. *Psychological Science*, *20*(9), 1169-1174.
- Kaspar, K. (2013). A weighty matter: Heaviness influences the evaluation of disease severity, drug effectiveness, and side effects. *PLoS ONE*, *8*(11): e78307. doi:10.1371/journal.pone.0078307
- Kaspar, K., & Krull, J. (2013). Incidental haptic stimulation in the context of flirt behavior. *Journal of Nonverbal Behavior*, *37*(3), 165-173. doi:10.1007/s10919-013-0154-0
- Kouchaki, M., Gino, F., & Jami, A. (2014). The burden of guilt: Heavy backpacks, light snacks, and enhanced morality. *Journal of Experimental Psychology General*, *143*(1), 414-424. doi:http://dx.doi.org/10.1037/a0031769
- Lakens, D. (2014). What p-hacking really looks like: A comment on Masicampo and LaLonde (2012). *Quarterly Journal of Experimental Psychology*. doi:10.1080/17470218.2014.982664

- Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical & Statistical Psychology*, *31*, 107-112.
- Maglio, S. J., & Trope, Y. (2012). Disembodiment: Abstract construal attenuates the influence of contextual bodily state in judgment. *Journal of Experimental Psychology General*, *141*(2), 211-216. doi:<http://dx.doi.org/10.1037/a0024520>
- Moreno, S. G., Sutton, A. J., Abrams, K. R., Cooper, N. J., Turner, E. H., Palmer, T. M., & Ades, A. E. (2009). Novel methods to deal with publication biases: Secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *BMJ*, *339*(b2981). doi:[10.1136/bmj.b2981](https://doi.org/10.1136/bmj.b2981)
- Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, *9*(2). doi:[10.1186/1471-2288-9-2](https://doi.org/10.1186/1471-2288-9-2)
- Nuijten, M. B., Hartgerink, C. H. J., Van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 1-22. doi: [10.3758/s13428-015-0664-2](https://doi.org/10.3758/s13428-015-0664-2)
- R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Rabelo, A. L. A., Keller, V. N., Pilati, R., & Wicherts, J. M. (2015). No effect of weight on judgments of importance in the moral domain and evidence of publication bias from a meta-analysis. *PLoS ONE*, *10*(8): e0134808. doi:[10.1371/journal.pone.0134808](https://doi.org/10.1371/journal.pone.0134808)
- Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (pp. 295-315). New York: Russell Sage Foundation.
- Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 97-128. doi:[10.1348/000711007X255327](https://doi.org/10.1348/000711007X255327)
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366. doi:[10.1177/0956797611417632](https://doi.org/10.1177/0956797611417632)

- Simonsohn, U. (2015, February 9). "The" effect size does not exist [Web log message]. Retrieved from <http://datacolada.org/2015/02/09/33-the-effect-size-does-not-exist/>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9(6), 666-681. doi:10.1177/1745691614553988
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2), 534-547. doi:10.1037/a0033242
- Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1), 60-78.
- Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, 49(1), 108-112. doi:10.2307/2684823
- Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99-110). Chichester, UK: Wiley.
- 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(6), 1137-1145. doi:10.1037/xge0000086
- Van Aert, R. C. M., Maassen, E., Wicherts, J. M., & Van Assen, M. A. L. M. (2016). Estimating effect size in meta-analysis in the presence of publication bias and questionable research practices. Manuscript in preparation.
- Van Assen, M. A. L. M., Van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, 20(3), 293-309. doi:<http://dx.doi.org/10.1037/met0000025>
- Viechtbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *Journal of Educational and Behavioral Statistics*, 32(1), 39-60. doi:10.3102/1076998606298034
- Witkovský, V. (2013). A note on computing extreme tail probabilities of the noncentral t-distribution with large noncentrality parameter. *Acta Universitatis Palackianae*

Olomucensis. Facultas Rerum Naturalium. Mathematica, 52(2), 131-143.
doi:<http://dml.cz/handle/10338.dmlcz/143545>

Zhang, M., & Li, X. (2012). From physical weight to psychological significance: The contribution of semantic activations. *Journal of Consumer Research*, 38(6), 1063-1075.
doi:10.1086/661768

Footnotes

¹ This principle is one of the most fundamental principles of probability and statistics. For instance, this principle is applied when sampling from distributions using so-called “inverse transform sampling” or the “inverse CDF method” (Gentle, 2004, Chapter 4). In this method, one starts sampling a random number from a uniform distribution from 0 to 1. Next, the random number, which is considered a cumulative percentage of the distribution, is used to calculate the *x*-value of the distribution that one wished to sample in the first place.

² The distribution of (conditional) *p*-values based on the true effect size is only uniform when the assumptions of the underlying statistical model (e.g. independence, effect distribution are valid [Bland, 2013]).

³ If *p*-hacking and publication bias were absent in these three studies, the fixed-effect meta-analytic estimator is unbiased and most efficient. In case publication bias was present but none of the three studies involved *p*-hacking, only the estimators of *p*-uniform and *p*-curve would have been accurate. Estimators of both fixed-effect meta-analysis and *p*-uniform and *p*-curve are inaccurate if *p*-hacking was used in the primary studies.

⁴ Functions for applying *p*-uniform can be loaded in R by means of running the following code: `devtools::install_github("RobbievanAert/puniform"); library(puniform)`

⁵ Due to transformation of d_{obs} and d_{cv} to *z*-values (see later on in this section), conditional *p*-values in *p*-uniform are divided by a value that is slightly larger than .05. Furthermore, dividing by .05 is only feasible if all observed effect sizes are statistically significant in the same direction. Imagine a situation where the observed effect size of Effect 1 is changed into $d = -0.872$, $t(48) = -3.133$. The two-tailed *p*-value of Effect 1 remains .00294, but

the observed effect size is statistically significant in the opposite direction than Effect 2 and 3. *P*-uniform and *p*-curve use one-tailed *p*-values and consequently, effects with opposite sign will be omitted when applying *p*-uniform or *p*-curve. If statistically significant effect sizes in a meta-analysis are observed in both tails of the distribution, it is advised to apply *p*-uniform and *p*-curve to both the statistically significant observed positive effect sizes and the statistically significant observed negative effect sizes. An example of such an analysis is described in Simonsohn et al. (2014a).

⁶ *P*-uniform's confidence interval is obtained by means of test inversion (e.g., Casella & Berger, 2002, Chapter 9), so the lower (upper) bound of the confidence interval equals that effect size for which the sum of conditional *p*-values is equal to the 2.5th (97.5th) percentile of the Irwin-Hall distribution. The statistical test of the null-hypothesis of no effect of *p*-uniform examines whether the conditional *p*-values follow a uniform distribution if $\delta=0$ (Van Assen et al., 2015).

For the publication bias test of *p*-uniform, all studies (significant and nonsignificant) in a meta-analysis are used for computing the effect size estimate based on fixed-effect meta-analysis. Only the statistically significant studies are then used to examine whether the conditional *p*-values follow a uniform distribution conditional on this fixed-effect meta-analytic effect size estimate. If the statistically significant *p*-values are not uniformly distributed conditional on this effect size estimate, the null-hypothesis of no publication bias is rejected (Van Assen et al., 2015).

⁷ For illustrative purposes we designed an example where *p*-uniform's effect size estimate equals the true effect size; just as easily an example can be constructed where *p*-curve's estimate equals the true effect size.

Appendix A

Illustration of logic of and computations in *p*-uniform and *p*-curve

A simple example involving one control condition and one experimental condition with 25 observations each illustrates the logic underlying *p*-uniform and *p*-curve. Imagine that the true effect size equals $\delta=0.5$ and three statistically significant effects were observed using a two-tailed *t*-test ($\alpha = .05$) for testing the null-hypothesis of no effect; Effect 1: $d = 0.872$, $t(48) = 3.133$ (two-tailed $p = .00294$); Effect 2: $d = 0.737$, $t(48) = 2.302$ (two-tailed $p = .0110$); and Effect 3: $d = 0.641$, $t(48) = 2.302$ (two-tailed $p = .0257$). Applying traditional fixed-effect meta-analysis to these three observed studies yields an overestimated effect size of 0.748 (95% CI: 0.416; 1.080).

P-curve and *p*-uniform employ the conditional *p*-values, that is, conditional on the effect size being statistically significant. More precisely, the conditional *p*-value of an observed effect size refers to the probability of observing this effect size or larger, conditional on the observed effect size being statistically significant *and* given a particular population (or “true”) effect size. Statistical significance has to be taken into account because *p*-uniform and *p*-curve only focus on the interval with *p*-values between 0 and .05 rather than the interval from 0 to 1. Figure A1 depicts how this conditional *p*-value of Effect 3 is computed for three different candidates of the underlying effect size, namely $\delta=0$, $\delta=0.5$ (i.e., the true effect size), and $\delta=0.748$ (i.e., estimate of fixed-effect meta-analysis). Figure A1a reflects the conditional *p*-value for $\delta=0$, which is calculated by dividing the probability of observing an effect size larger than the observed Effect 3 (dark grey area in Figure A1a to the right of d_{obs}) by the probability of observing an effect size larger than the critical value (light and dark grey area in Figure A1a to the right of d_{cv}). For $\delta=0$, the null-hypothesis being tested, this boils down to dividing the *p*-value (.0257) by $\alpha=.05$,

yielding a conditional p -value (denoted by q) for Effect 3 of $q_3 = .0257 / .05 = .507$.⁵ Thus, for $\delta = 0$ the conditional p -value is simply 20 times the traditional p -value.

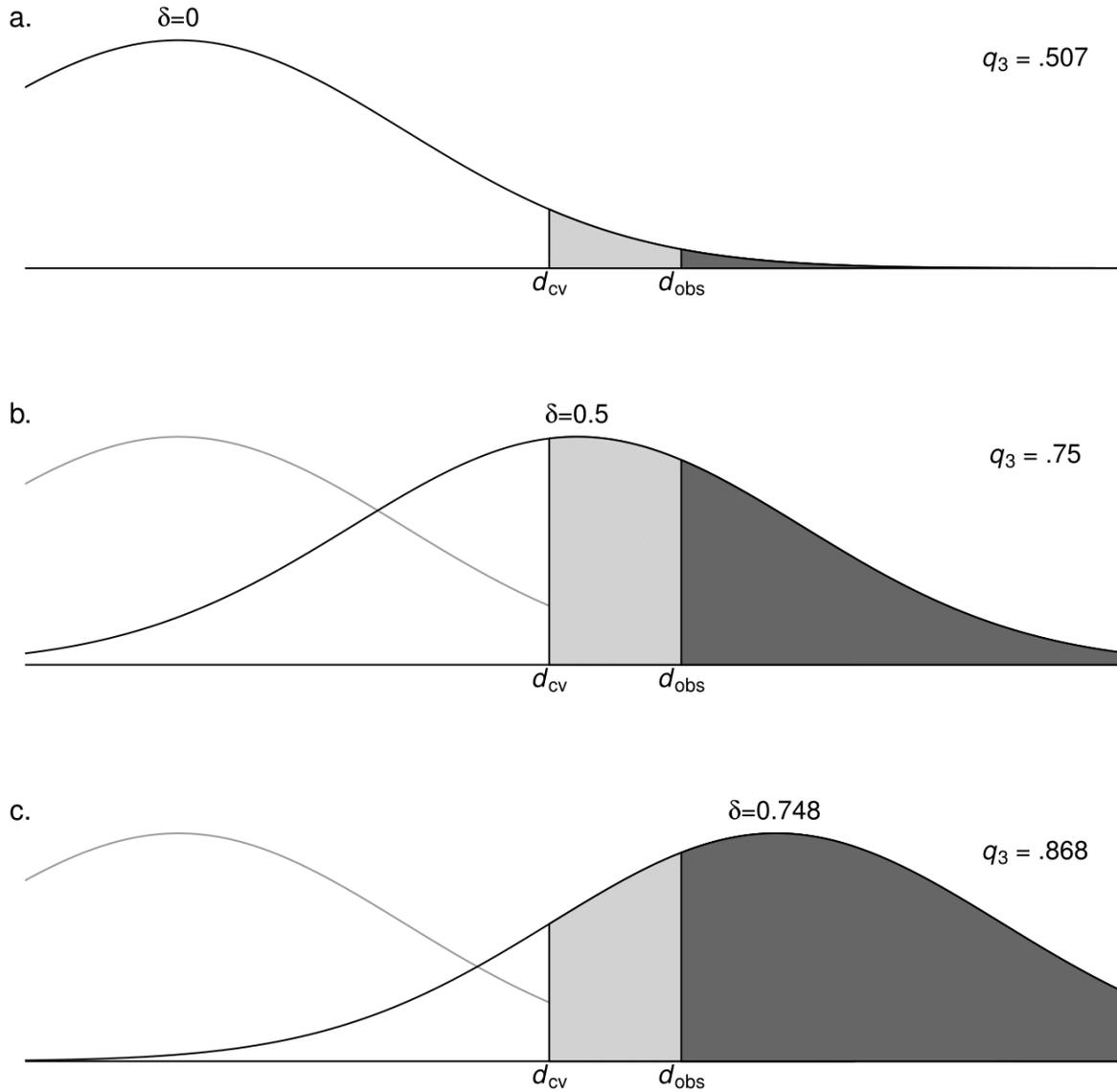


Figure A1. Illustration of computation of conditional p -values for Effect 3 (q_3) for three effect sizes: a. $\delta = 0$; b. $\delta = 0.5$ (true effect size); c. $\delta = 0.748$ (estimate of fixed-effect meta-analysis).

Critical value on Cohen's d scale is denoted by d_{cv} and observed effect size is denoted by d_{obs} .

Computation of the conditional p -values under effects that differ from zero uses calculations closely resembling the computation of statistical power of a test. Consider the conditional p -value of Effect 3 at $\delta=0.5$ (Figure A1b). The critical value (d_{cv}) and the observed effect size (d_{obs}) on the Cohen's d scale remain the same, but the distribution of true effect size δ is now shifted to the right. The numerator in computing the conditional p -value expresses the probability that the observed effect size d_{obs} is 0.641 or larger given $\delta=0.5$ (dark grey area in Figure A1b to the right of d_{obs}), which equals 0.314, whereas the denominator expresses the probability that the observed effect size is statistically significant given $\delta=0.5$ (light and dark grey area in Figure A1b to the right of d_{cv}), which equals 0.419 (i.e., the traditional power of the study given its degrees of freedom and $\delta=0.5$). This yields a conditional p -value for Effect 3 at $\delta=0.5$ of $q_3=0.314/0.419=0.75$. The conditional p -value of Effect 3 at $\delta=0.748$, as displayed in Figure A1c, can be computed in a similar way: $q_3=0.644/0.742=0.868$.

The conditional p -values of all three observed effect sizes in our example under the three different true effect sizes are presented in Figure A2. The solid black lines in the left panel of Figure A2 shows the conditional p -values for $\delta=0$:

$$q_1=.00294/.05=.0558 \quad q_2=.0110/.05=.213 \quad q_3=.0257/.05=.507.$$

The dashed grey lines in the left panel illustrate uniformly distributed conditional p -values. In case of three studies these uniformly distributed conditional p -values should equal $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{3}{4}$. Note that the observed conditional p -values, summing to $.0558+.213+.507=.776$, are lower than their corresponding expected uniformly distributed conditional p -values, which sum to $\frac{1}{4}+\frac{1}{2}+\frac{3}{4}=1.5$. Hence, we see that the conditional p -values under the null hypothesis ($\delta=0$) as given in the left-hand side of Figure A2 do *not* fit a uniform distribution.

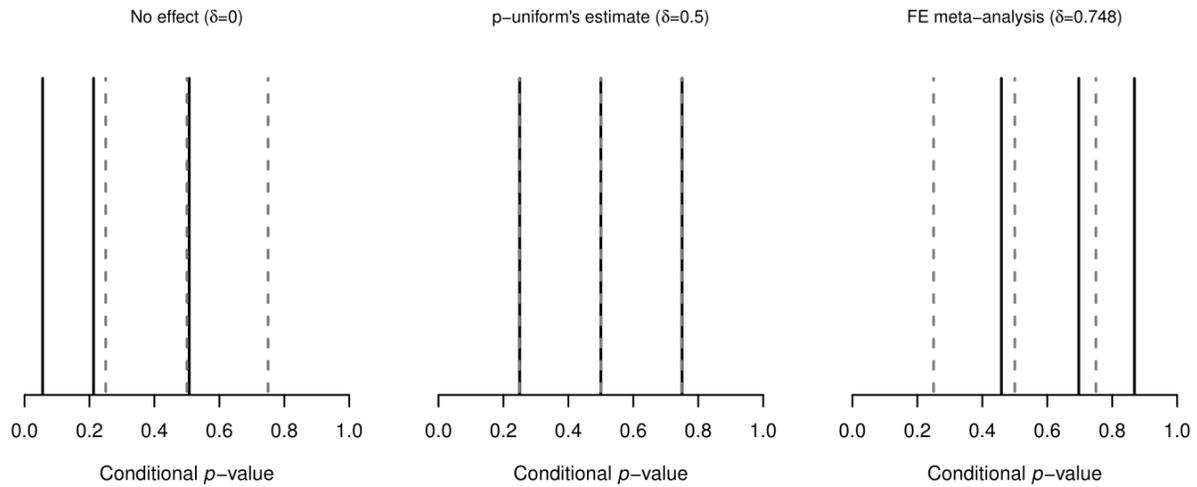


Figure A2. Observed conditional *p*-values (solid black lines) and conditional *p*-values under uniformity (dashed gray lines) for the example with three observed effect sizes. The three panels refer to the conditional *p*-values for *p*-uniform's hypothesis test of no effect ($\delta=0$), *p*-uniform's effect size estimate ($\delta=0.5$), and effect size obtained by fixed-effect meta-analysis ($\delta=0.748$).

To obtain the effect size estimate of *p*-uniform, effect size (δ) has to be shifted until the sum of conditional *p*-values equals 1.5, which is the expected value of the sum under uniformity, i.e. given the true effect size. Figure A3 shows the effect of shifting δ on the conditional *p*-values from -.5 to 1.5 for the three observed effect sizes in our example. Each conditional *p*-value increases when the true effect size gets larger. For instance, the conditional *p*-value of Effect 1 increases from .0558 to .25 when the true effect size is increased from 0 to .5, and further increased to .459 if true effect size is increased to .748. As a consequence of these increases, the sum of conditional *p*-values also increases as true effect size increases.

The middle panel in Figure A2 presents the conditional *p*-values in case the effect size is

shifted to $\delta=0.5$. These conditional *p*-values are also shown in Figure A3 as the intersections of the three curves with the vertical line representing $\delta=0.5$, and equal:

$$q_1=.25 \quad q_2=.50 \quad q_3=.75.$$

These conditional *p*-values exactly match (and studies were selected to exactly match) the expected conditional *p*-values under uniformity. Consequently, the sum of the conditional *p*-values also equals the sum of the conditional *p*-values under uniformity (1.5). This indicates that the effect size estimate of *p*-uniform will be equal to the true effect size of 0.5.

The right panel in Figure A2 and the intersections of the studies' curves with line $\delta=0.748$ in Figure A3 show the conditional *p*-values conditional on the effect size $\delta=0.748$, which was the estimate of traditional fixed-effect meta-analysis:

$$q_1=.459 \quad q_2=.697 \quad q_3=.868.$$

All are higher than their corresponding expected conditional *p*-values under uniformity, and their sum (2.031) is larger than the expected sum under uniformity (1.5). These results indicate that traditional fixed-effects meta-analysis overestimated the effect size. If this occurs, it is not farfetched to suppose that publication bias exists, i.e. some nonsignificant results are missing from the set of studies included in the meta-analysis.

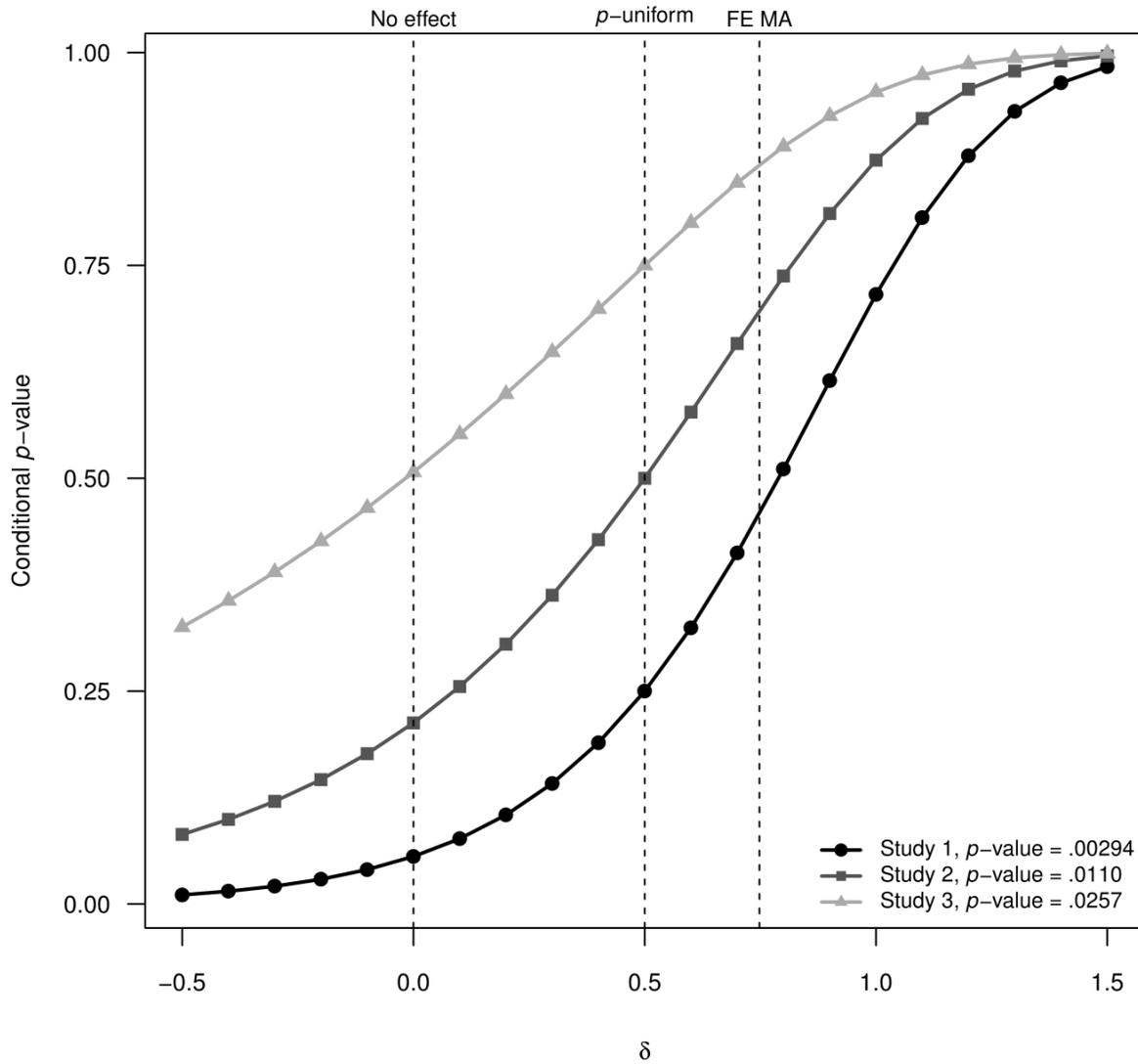


Figure A3. Conditional *p*-values as a function of true effect size (*x*-axis) for each of the three observed effect sizes in the example. Effect sizes zero, true effect size, and estimated by fixed-effects meta-analysis are indicated by vertical lines.

Table A1 shows the results of applying *p*-uniform and *p*-curve to the example. The

estimated effect size by *p*-uniform is exactly equal to the true effect size of $\delta = 0.5$. Other output of *p*-uniform is the 95% confidence interval (-0.300; 0.960), and that both the null-hypothesis of no effect ($p = .0737$) and the hypothesis of no publication bias ($p = .147$) cannot be rejected.⁶ The output of *p*-curve incorporate neither a confidence interval nor a publication bias test. *P*-curve's estimate of .511 is slightly larger than the true effect size⁷, and *p*-curve's result of the test of no effect is $p = .086$. Why are the results of both methods different, if they are based on the same logic? This is because the methods differ in implementation, which we explain in the supplementary materials.

Table A1. Results of *p*-uniform and *p*-curve when applied to the artificial example based on three observed effect sizes, with $\delta=0.5$.

	<i>p</i> -uniform	<i>p</i> -curve
Effect size estimate	0.500	0.530
95% CI	(-0.308;0.964)	-
Test of $H_0: \delta = 0$	$z=-1.44; p=.0753$	$\chi^2(6)=1.97; p=.0772$
Publication bias test	$z=1.06; p=.144$	-

Appendix B

Table B1. Studies and corresponding sample sizes (group 1: n_i^1 and group 2: n_i^2), *t*-values and two-tailed *p*-values as included in the meta-analysis described in Rabelo, et al. (2015).

Study no.	Article and experiment	n_i^1	n_i^2	<i>t</i> -value	<i>p</i> -value (two-tailed)
1	Ackerman et al. (2010), Exp. 1	26	28	2.016	0.0489
2	Ackerman et al. (2010), Exp. 2	21	22	1.867	0.0690
3	Chandler et al. (2012), Exp. 2	30	30	2.554	0.0133
4	Chandler et al. (2012), Exp. 1	50	50	2.113	0.0372
5	Chandler et al. (2012), Exp. 3	50	50	2.390	0.0188
6	Hafner (2013), Exp. 1	30	30	2.042	0.0457
7	Jostmann et al. (2009), Exp. 1	20	20	2.245	0.0307
8	Jostmann et al. (2009), Exp. 2	22	28	2.081	0.0428
9	Jostmann et al. (2009), Exp. 3	25	24	2.191	0.0335
10	Jostmann et al. (2009), Exp. 4	20	20	2.294	0.0274
11	Kaspar & Krull (2013)	45	45	3.049	0.0030
12	Kouchaki et al. (2014), Exp. 1a	15	15	2.020	0.0531
13	Kouchaki et al. (2014), Exp. 1c	27	27	2.184	0.0335
14	Kouchaki et al. (2014), Exp. 2	26	25	2.307	0.0254
15	Kouchaki et al. (2014), Exp. 3	35	36	2.308	0.0240
16	Kaspar (2013), Exp. 1	20	20	3.268	0.0023
17	Kaspar (2013), Exp. 2	25.5	25.5	2.306	0.0254
18	Kaspar (2013), Exp. 3	31	31	2.278	0.0263

19	Kaspar (2013), Exp. 4	48.5	48.5	2.053	0.0429
20	Kaspar (2013), Exp. 5	30	30	2.452	0.0172
21	Kouchaki et al. (2014), Exp. 4	31	31	2.139	0.0365
22	Maglio and Trope (2012), Exp. 2	18	18	2.284	0.0287
23	Zhang and Li (2012), Exp. 1	35	35	2.382	0.0200
24	Zhang and Li (2012), Exp. 2	39	39	1.994	0.0498
25	Zhang and Li (2012), Exp. 4	40	40	2.530	0.0134

Supplementary materials

Conducting meta-analyses based on p -values: Reservations and recommendations for applying p -uniform and p -curve

Robbie C. M. van Aert, Jelte M. Wicherts, and Marcel A. L. M. van Assen

Table of Contents:

1. Technical details on differences in implementation between p -uniform and p -curve
2. Explanation of p -uniform's highly negative estimates and p -curve's erratic behavior
3. Details of simulations p -hacking

1. Technical details on differences in implementation between *p*-uniform and *p*-curve

P-uniform and *p*-curve differ in three important ways with respect to how the null-hypothesis of no effect is tested and the effect size is estimated. First, *p*-uniform's hypothesis test of no effect is based on a left-tailed *z*-test,

$$z = \frac{\sum_{i=1}^K q_i^0 - K \times 0.5}{\sqrt{K/12}},$$

where K is the number of statistically significant studies in a meta-analysis, $\sum_{i=1}^K q_i^0$ is the sum of the conditional *p*-values given an effect size of zero and $K \times 0.5$, and $\sqrt{K/12}$ are the mean and standard deviation of the uniform distribution, respectively. In contrast, *p*-curve uses Fisher's method (Fisher, 1925, Chapter 4) for the test of no effect. The test statistic of Fisher's method is a χ^2 value with $2K$ degrees of freedom and is computed by $-2 \sum_{i=1}^K \ln(q_i^0)$.

A second difference in implementation is that *p*-curve computes conditional *p*-values via the *t*-distribution whereas *p*-uniform transforms the effect sizes to *z*-values via Hedges' *g* before computing the conditional *p*-values. Transforming to *z*-values is necessary because computing conditional *p*-values based on the *t*-distribution yields inaccurate conditional probabilities for effects with *p*-values close to .05, or even undefined conditional *p*-values (due to division by zero). The section in the article on the sensitivity of both *p*-uniform and *p*-curve to *p*-values close to .05 describes the consequences of this inaccuracy on *p*-curve's effect size estimation (see also Figure 2).

Third, *p*-curve minimizes the Kolmogorov-Smirnov test statistic to obtain an effect size estimate, whereas the current default implementation of *p*-uniform uses the Irwin-Hall distribution. The Irwin-Hall distribution is the distribution of the sum of independent uniformly distributed random variables. The Kolmogorov-Smirnov test statistic and estimation using the

Irwin-Hall distribution differ in how the test statistics are computed and how they are optimized.

Using the Irwin-Hall distribution, *p*-uniform estimates effect size such that the sum of the conditional *p*-values equals $0.5K$. The lower (upper) bound of the confidence interval equals that effect size for which the sum of conditional *p*-values equals the 2.5th (97.5th) percentile of the Irwin-Hall distribution. For instance, the Kolmogorov-Smirnov statistic assesses misfit of the conditional *p*-value distribution with respect to the uniform distribution. The estimate based on the Irwin-Hall distribution equals that value for which the sum of conditional *p*-values equals the expected value of this distribution (which equals half the number of statistically significant studies included in the analysis).

Other methods exist that can be used to examine whether a distribution of conditional *p*-values follow a uniform distribution and can, therefore, also be used for estimation of effect size. Fisher's method (1925, Chapter 4) can also be used for estimating effect size and is implemented in *p*-uniform by $-\sum_{i=1}^K \ln(q_i)$ where q_i are the conditional *p*-values and follows a gamma distribution with K and 1 degrees of freedom. A slight variation of the Fisher's method is $-\sum_{i=1}^K \ln(1 - q_i)$ and has the property that it weights studies with high conditional *p*-values (i.e., studies with *p*-values just below .05) more heavily than the original Fisher's method. A confidence interval based on both Fisher's methods equals that effect size for which the method's test statistic is equal to the 2.5th and 97.5th percentile of the gamma distribution with K and 1 degrees of freedom. Another method that can be used for effect size estimation is the Anderson-Darling test (Anderson & Darling, 1954). All five aforementioned estimators are currently optional choices in *p*-uniform.

The two estimators based on Fisher's method ("*p*" and "*1-p*") were used in *p*-uniform's implementation in Van Assen et al. (2015), but the Irwin-Hall estimator is now the default

because this estimator has several desirable properties. First, its math, summing *p*-values, and the intuition underlying the estimator are in our opinion easier to grasp than of the other estimators. Second, as a consequence, it underlies our recommendation (Table 1) that the estimated effect is smaller, equal, larger than 0 if the average of included *p*-values is larger, equal, smaller than 0.025, respectively. Third, it is rather straightforward to compute confidence intervals using the Irwin-Hall estimator, which is not the case using the Kolmogorov-Smirnov and Anderson-Darling estimator. And finally, it is relatively robust to outliers, i.e., studies with an outlying effect size or very large sample size, compared to the methods based on Fisher's method.

2. Explanation of *p*-uniform's highly negative estimates and *p*-curve's erratic behavior

As described in the section “Sensitivity to *P*-values Close to .05”, *p*-uniform yields implausible highly negative effect size estimates if the majority of *p*-values in a meta-analysis is just below .05. This can be explained by conditional *p*-values which are less sensitive to changes of effect size for observed effect sizes that are just statistically significant. Figure S1 presents the conditional *p*-values as a function of effect size and the *p*-value of a study based on a two-group design with 50 participants per group. The curve labeled “ $p = 0.025$ ” corresponds to the conditional *p*-value of an observed effect size that, when the hypothesis of no effect is tested, yields a *p*-value of 0.025. Note that this conditional *p*-value equals .5, or the expected value, for effect size equal to 0 (.025 being the midpoint of the interval [0, .05]). The figure illustrates that a curve's slope gets smaller when the *p*-value gets closer to .05. This may have dramatic implications on the estimation of effect size, as in our example described in the article. We had three studies (Effect 1 with $d = .963$, $t(18) = 2.154$, $p = .045$ (two-tailed), Effect 2 with $d = .582$, $t(48) = 2.058$, $p = .045$, and Effect 3 with $d = .4$, $t(98) = 2.002$, $p = .048$), meaning that their conditional *p*-values are close to 1 for sensible effect size values. Hence, to obtain an *average* conditional *p*-value of .5, effect size needs to be shifted much to the left (even outside the figure).

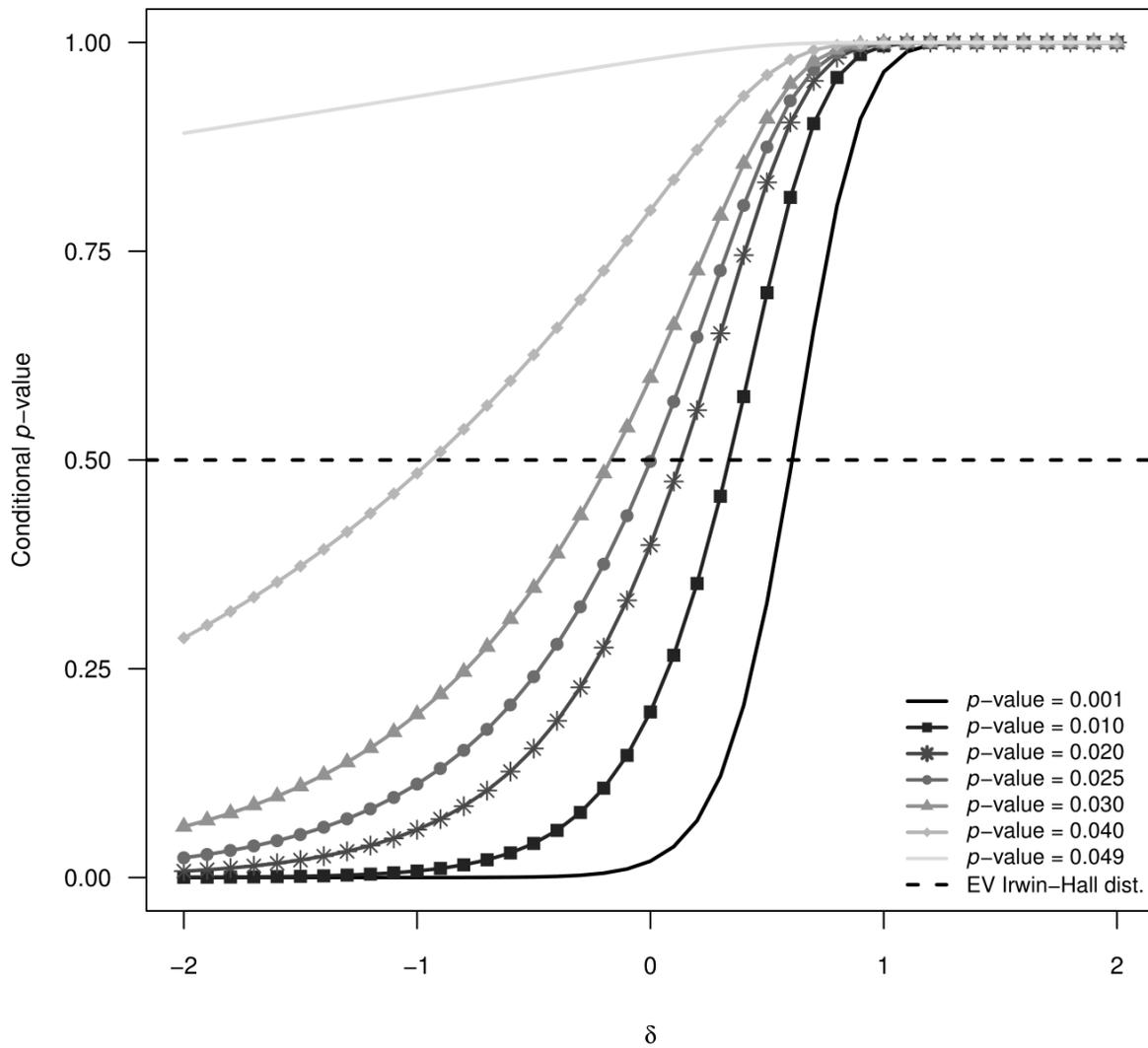


Figure S1. Conditional *p*-values as a function of effect size (*x*-axis) for a selection of one-tailed *p*-values (solid lines) based on a two-group design with 50 participants per group. The dashed black line refers to the expected value of the Irwin-Hall distribution in case of one study.

3. Details of simulations *p*-hacking

For "First significant DV" we used the same procedure as Simonsohn et al. (2014a); each study had two conditions with 20 observations and means were compared using a *t*-test on three dependent variables A, B, C, correlated with each other at $r = .49$. If the effect on A was statistically significant, then it was reported and the results of B and C were not reported. Otherwise the result of B was considered; if B yielded a statistically significant outcome, then only that result was reported; if the effects on neither A nor B were statistically significant, the outcome on C was reported, but only if that effect was statistically significant. If none of the effects on A, B, and C were significant, the dependent variable that provided the smallest *p*-value was the one being reported and included in the meta-analysis. The same procedure was followed for "DV with lowest *p*-value", but the dependent variable with the lowest *p*-value was reported instead of the first significant dependent variable.

In "Multiple conditions", one control and three independent but identical (i.e., same population mean) experimental conditions were run, and each experimental condition was compared to the control using a *t*-test. Only the result of the most significant test (lowest *p*-value) was reported. Finally, in data peeking (a.k.a. optional stopping) means of two groups with 20 observations each were compared using a *t*-test. If a statistically significant finding was obtained, the study was 'published', otherwise seven participants per group were added to the data and another hypothesis test was conducted. This procedure was repeated until a statistically significant effect was found or the procedure was applied three times.

References

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268), 765-769.

Fisher, R. A. (1925). *Statistical methods for research workers* (1st ed.). London: Oliver & Boyd.