# Publication Bias and Psychotherapy: A Pre-analysis Plan[*]

Johannes Haushofer[†] Daniel Mellow[†] Pim Cuijpers[‡]

November 8, 2017

**Corresponding author:**
Johannes Haushofer, Ph.D.
Assistant Professor
Princeton University
Department of Psychology
427 Peretsman Scully Hall
Princeton, NJ 08544
Phone: 617-360-1605
haushofer@princeton.edu

[*] We thank Robbie van Aert, Daniel Benjamin, Blakeley McShane, and seminar participants at the Busara Center for Behavioral Economics for helpful comments and suggestions, and Robbie van Aert, Isaiah Andrews, Blakeley McShane and Felix Schönbrodt for providing code.

[†] Princeton University, Princeton, NJ, USA, and Busara Center for Behavioral Economics, Nairobi, Kenya

[‡] VU Amsterdam, Amsterdam, The Netherlands

**Abstract**

This document outlines the pre-analysis plan (PAP) for a meta-analysis of randomized control trials (RCTs) which investigate the effectiveness of psychotherapies for treating depression. We will analyze a database of psychotherapy RCTs using a set of techniques to test for and correct for publication bias, estimate power, and calculate pre- and post-experimental odds. This plan describes the analysis methods that will be used.

Keywords: Cognitive-Behavioral Therapy, Meta-analysis Publication Bias, Questionable Research Practices

# 1. Introduction

Unipolar major depressive disorder (MDD) is one of the most prevalent mental disorders worldwide. The standard therapeutic approach is psychotherapy, which was first developed in the 19th century and has experienced a surge in usage since the 1960s. Currently, the most prominent approach to treating depression is cognitive-behavioral therapy (CBT), which focuses on correcting distorted thinking. More recently, other therapies such as interpersonal therapy and problem-solving therapy have gained prominence. Psychotherapies for depression in general, and CBT in particular, are widely thought to be well-supported by empirical evidence; so much so that research efforts are now focused not on validating psychotherapy against control, but on comparing the effectiveness of different forms of psychotherapy to each other [Barth et al., 2013], or comparing it to medication [Amick et al., 2015]. Existing meta-analyses comparing psychotherapy against control conditions argue that it is effective in treating depression [Khan et al., 2012, Cuijpers et al., 2008].

How solid is the empirical evidence on which the widespread acceptance and usage of psychotherapy rest? We propose to analyze the results of all randomized experiments that have tested the effectiveness of the seven main types of psychotherapy against control conditions in adults. Typically these studies delivered treatment to individuals who were initially diagnoses as depressed for 8–12 weeks, and then measured depression symptoms. The control conditions were either a waitlist condition, a placebo drug, or "usual care", i.e. referral to mental health services in the community. We use meta-analytic approaches including recently developed statistical tools to assess and correct for publication bias, estimate power, and calculate pre- and post-experimental odds.

# 2. Data

We obtain data from a database of all randomized controlled trials comparing psychotherapy to a control condition, initially compiled by Cuijpers et al. [2008] and recently updated as of January 1, 2017. More detailed information can found at www.evidencebasedpsychotherapies.org. The data includes all randomized controlled trials studying the effect of psychotherapy on depression that contained a control group (either waitlist, placebo, or usual care).

# 3. Estimation Methods

## 3.1 P-curve and P-uniform

We will apply and present the results of the p-uniform method [van Assen et al., 2015]. This approach is based on the fact that *if the null hypothesis is true* the p-values of hypothesis tests follow a uniform distribution. Therefore a true mean of a sampling distribution can be estimated by finding the mean which produces a distribution of conditional p-values that is closest to the uniform distribution. Measuring the "closeness" of an empirical and ideal distribution can be done several ways:

1. Simonsohn et al. [2014b] minimize the Kolmogorov-Smirnov statistic, equal to the maximum distance between an empirical and ideal cumulative distribution function. This special case of the p-uniform approach is known as the "p-curve" method.

2. van Assen et al. [2015] calculate the Fisher statistic, equal to the sum of negative logarithms of the conditional p-values. If the distribution of p-values were uniform, this statistic would follow the gamma distribution, with a shape parameter equal to the number of studies minus two, and a scale parameter of one. Therefore, their effect size estimate is that which produces a Fisher statistic closest to the mean of this gamma distribution. Because probabilities are bounded by 0 and 1, the authors recommend an estimator which is actually based on the *complements* of the p-values. We will calculate both.

3. van Aert et al. [2016] update their previous p-uniform estimator by noting that the sum of standard uniform random variables follows the Irwin-Hall distribution. Therefore, they calculate the sum of conditional p-values and compare it to the mean of the Irwin Hall distribution.

We will report all four of these estimates (including two types of Fisher statistic), both for the sample as a whole and separately for each type of psychotherapy.

We will also report the results of a test for publication bias based on the "Fisher statistic" method described in (2) above. Note that the null hypothesis that there is no publication bias in a set of studies is equivalent to the claim that the mean effect size of the studies is an unbiased estimator of the true effect size. This statement can be tested by calculating a p-value for each study of the hypothesis that the true effect is the mean value among the studies. Then under the null hypothesis of no publication bias

$$L^{\bar{\mu}} = -\sum_{k=1}^{K} ln(1 - q_k^{\bar{\mu}}) \sim \Gamma(K-2, 1)$$

Where $\bar{\mu}$ is the mean estimated effect size, $q_k^x$ represents the p-value of study $k$ for effect size $x$, and $L^x$ is the Fisher statistic. We reject the null hypothesis of no publication bias if the Fisher statistic is above the $1 - \frac{\alpha}{2}$ quantile or below the $\frac{\alpha}{2}$ quantile of the Gamma distribution [Simonsohn et al., 2014a,b, van Assen et al., 2015, van Aert et al., 2016]. Note, however, that this test relies on the same assumptions as the estimator, namely homogeneity of effect size and researcher honesty.

## 3.2 Parametric Selection Models

We will implement two models which attempt to explicitly model the publication bias process. The three-parameter selection model [PSM; Carter et al., 2017, McShane et al., 2016] approach involves making a distributional assumption about the underlying data-generating process of studies and a specific selection function for publication bias. Doing so allows for derivation of a probability distribution function of observed studies. The parameters of that distribution are then estimated through maximum likelihood. In practice, this is done by assuming that the sampling distribution of studies is normally distributed and that publication probability is a step function of the p-value, with discontinuity at $p = 0.05$. If effect sizes are standardized, this model is characterized by three parameters:

1. The mean true effect size.

2. The ratio of the probability of publication of nonsignificant studies to that of statistically significant studies.

3. A heterogeneity parameter which captures the variation in the the underlying true effect size measured by each study.

McShane et al. [2016] point out that a restricted version of the model, assuming complete publication bias (no insignificant studies published) and no heterogeneity, such that there is only one free parameter, is isomorphic to the theoretical foundation of p-curve/p-uniform. In this case the only difference is in identification strategy: p-curve/p-uniform essentially engages in moment matching, while the 1PSM utilizes a maximum likelihood estimator. Since maximum likelihood estimators are asymptotically efficient, McShane et al. [2016] argue (and demonstrate through simulations) that the 3PSM is superior in a wide variety of settings. The relative difference in estimator performance is small, though, and under debate at the time of writing [Nelson et al., 2017].

We will implement the 3PSM using the custom R function available in the supplementary material of McShane et al. [2016]. We will report the three parameter estimates for the sample as a whole and separately for each type of psychotherapy.

Secondly, we will apply another parametric selection model described by Andrews and Kasy [A-K; 2017]. A-K set up a very general framework for analyzing the data generating process of observed studies under publication bias. In practice, however, they make the same distributional and functional assumptions as McShane et al. [2016] to identify parameters of interest. That is, publication bias is assumed be to a step function and the distribution of effect sizes to be normal. In addition, A-K assume that effect sizes and sample sizes are independent. While this assumption is common — and the foundation for older, "funnel-plot" tests of publication bias — it is criticized as unrealistic, given that researchers in practice place great emphasis on power calculations for determining sample size [Lau et al., 2006].

Making these identifying assumptions allows A-K to derive the cumulative distribution function of observed effect sizes, and therefore moment-matching estimators for mean effect size, publication bias, and heterogeneity. We will implement the A-K estimator using a custom R package, available from the authors on request.

# 4. Post-estimation Analysis

Following application of the methods detailed in sections 3.1 and 3.2, we will generate six preferred estimates of the true effect of cognitive-behavioral therapy on major depression. We decide not to choose between these estimates. Instead, we illustrate the implications of those estimates, along with the fixed and random effects meta-analytic estimates, for the literature.

## 4.1 Test of Null Effect

For our naive meta-analytic estimate, statistical inference is straightforward. For the publication bias-corrected estimates, however, constructing valid confidence sets is difficult and subject to assumptions about the data-generating process of effect sizes. In particular, if there is underlying heterogeneity in the true effects that studies are estimating, then the calculated standard errors of the p-curve/p-uniform (Fisher Statistic and Irwin Hall) and Andrews-Kasy estimators will be lower bounds.

We will address this problem by simulating the distributions of these estimators under the null hypothesis that the true mean of the effect size distribution is zero. Under our simulation framework, these distributions will be conditioned on: (i) the degree of publication bias, assumed to be a step function with a discontinuity at statistical significance; (ii) heterogeneity, or the width of the distribution which generates the true effect size; and (iii) how aggressively authors engage in "Questionable Research Practices" (QRPs), sometimes referred to as "p-hacking."

Let $\beta$ represent the relative publication probability of a nonsignificant study and the true effect size underlying each study be drawn from a $N\big(0, \tau^2\big)$ distribution. Furthermore, consider four forms of QRPs: selection between two dependent variables; optional use of a covariate; optional stopping of data collection; and optional removal of outliers. Carter et al. [2017] develop an algorithmic framework for simulating these behaviors in the creation of meta-analytic samples. Parametrizing the algorithm is complicated, but we consider the limits of the spectrum that the authors develop: a scenario in which no researchers engage in QRPs, and a "high" QRP environment in which 50% of researchers use all four tactics, 40% use the first two, and 10% do not use any.

Therefore we define a set of parameter combinations that characterize a simulated data-generating process. We generate the eight distributions for each estimator, characterized by the corners of this set:

1. No publication bias ($\beta = 1$), no heterogeneity ($\tau = 0$), and no QRPs.

2. Complete publication bias ($\beta = 0$), no heterogeneity ($\tau = 0$), and no

QRPs.

3. No publication bias ($\beta = 1$), severe heterogeneity ($\tau = 0.5$), and no QRPs.
4. No publication bias ($\beta = 1$), no heterogeneity ($\tau = 0$), and a "high" QRP environment .
5. Complete publication bias ($\beta = 0$), severe heterogeneity ($\tau = 0.5$), and no QRPs.
6. Complete publication bias ($\beta = 0$), no heterogeneity ($\tau = 0$), and a "high" QRP environment.
7. No publication bias ($\beta = 1$), severe heterogeneity ($\tau = 0.5$), and a "high" QRP environment.
8. Complete publication bias ($\beta = 0$) , severe heterogeneity ($\tau = 0.5$), and a "high" QRP environment.

Note that the p-curve/p-uniform estimators take do not consider insignificant studies and will be unaffected by the differences in publication bias functions. For each of these distributions we will report the mean and standard error, along with the implied p-values of the null hypothesis given the observed estimates. We make no further inferential judgements about the "true" effect size.

## 4.2 Power Analysis

Several authors have considered ways to analyze post-hoc power of published studies as a proxy for the degree to which the extant literature is distorted by publication bias [Button et al., 2013, Ioannidis and Trikalinos, 2007, Schimmack et al., 2017]. We follow an approach which is most similar in spirit to Button et al. [2013]: for each of the eight candidate true effects (the fixed and random effects meta-analytic estimates, and the six corrected estimates), we will calculate the median post-hoc power in our sample conditional on that estimate.

## 4.3 Bayesian Statistics and the Rejection Ratio

A relatively recent body of literature attempts to develop Bayesian concepts for the predictive value of hypothesis tests. We focus on the two types of "rejection ratio" described by Bayarri et al. [2016]. The "pre-experimental" rejection ratio is defined as the ratio of the probability that a false null is rejected (otherwise known as the power of the test) to the probability that a true null is rejected (equal to the significance level for a valid test), i.e. the ratio of rejection probabilities under $H_1$ and $H_0$ , respectively. We will calculate and report rejection ratios based on post hoc power conditional on eight effect sizes: the fixed and random effects meta-analytic estimates and six corrected estimates.

The "post-experimental" rejection ratio is similar in spirit but takes into account the actual data. This statistic, $R_{post}$, is equivalent to the Bayes' factor for a test, i.e. the likelihood of the observed data under $H_1$ over the likelihood of the observed data under $H_0$. The Bayes' factor depends on the null hypothesis, expressed here as a prior distribution, for which a wide class of models are available. We will calculate four post-experimental rejection ratios for hypotheses about the true underlying effect size of psychotherapy on depression for each estimator.

1. A rejection ratio $R^U$ based on a prior distribution that is uniform between 0 and the estimated effect size, which is the least conservative prior that is nonincreasing away from the null hypothesis. In the case where there is no reason to believe the intervention will work based on prior evidence, it is natural to make a prior which is nonincreasing away from the null.

2. A rejection ratio $R^P$ based on the prior distribution which is a point mass at the estimated effect size, which is most favorable to the alternative hypothesis among all priors.

3. We will calculate a rejection ratio $R^M$ which is the result of power considerations by using a prior distribution which a single point of unit mass at the minimum detectable effect size (MDES) with 80% power. That

is, the effect size for which 80% of the conditional null sampling distribution is above the critical value of a hypothesis test of significance level 0.05. Assuming normality of errors in hypothesis testing, this value is approximately 2.8 times the standard error of the mean.

4. A rejection ration $R^N$ of the prior distribution that the data are normally distributed with variance equal to one around the estimated effect size.

Using this approach we will report 32 Bayesian rejection ratios in total, presenting a picture of the Bayesian evidential value for psychotherapy that is sensitive to prior distribution assumptions and estimation strategies.

# References

Halle R. Amick, Gerald Gartlehner, Bradley N. Gaynes, Catherine Forneris, Gary N. Asher, Laura C. Morgan, Emmanuel Coker-Schwimmer, Erin Boland, Linda J. Lux, Susan Gaylord, Carla Bann, Christiane Barbara Pierl, and Kathleen N. Lohr. Comparative benefits and harms of second generation antidepressants and cognitive behavioral therapies in initial treatment of major depressive disorder: systematic review and meta-analysis. *BMJ (Clinical research ed.)*, 351:h6019, 2015. ISSN 1756-1833.

Isaiah Andrews and Maximilian Kasy. Identification of and Correction for Publication Bias. Working Paper 23298, National Bureau of Economic Research, March 2017. URL http://www.nber.org/papers/w23298. DOI: 10.3386/w23298.

Jürgen Barth, Thomas Munder, Heike Gerger, Eveline Nüesch, Sven Trelle, Hansjörg Znoj, Peter Jüni, and Pim Cuijpers. Comparative Efficacy of Seven Psychotherapeutic Interventions for Patients with Depression: A Network Meta-Analysis. *PLOS Med*, 10(5):e1001454, May 2013. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001454. URL http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001454.

MJ Bayarri, Daniel J Benjamin, James O Berger, and Thomas M Sellke. Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72:90–103, 2016.

Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

Evan Carter, Felix Schönbrodt, Will M Gervais, and Joseph Hilgard. Correcting for bias in psychology: A comparison of meta-analytic methods. Working Paper, 2017.

Pim Cuijpers, Annemieke van Straten, Lisanne Warmerdam, and Gerhard Andersson. Psychological treatment of depression: A meta-analytic database of randomized studies. *BMC Psychiatry*, 8:36, 2008. ISSN 1471-244X. doi: 10.1186/1471-244X-8-36. URL http://dx.doi.org/10.1186/1471-244X-8-36.

John PA Ioannidis and Thomas A Trikalinos. An exploratory test for an excess of significant findings. *Clinical trials*, 4(3):245–253, 2007.

Arif Khan, James Faucett, Pesach Lichtenberg, Irving Kirsch, and Walter A. Brown. A systematic review of comparative efficacy of treatments and controls for depression. *PloS One*, 7(7):e41778, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0041778.

Joseph Lau, John PA Ioannidis, Norma Terrin, Christopher H Schmid, and Ingram Olkin. Evidence based medicine: The case of the misleading funnel plot. *BMJ: British Medical Journal*, 333(7568):597, 2006.

Blakeley B McShane, Ulf Böckenholt, and Karsten T Hansen. Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11(5):730–749, 2016.

Leif Nelson, Joseph Simmons, and Uri Simonsohn. [61] Why p-curve excludes ps>.05, June 2017. URL http://datacolada.org/61.

Ulrich Schimmack, Moritz Heene, and Kamini Kesavan. Reconstruction of a train wreck: How priming research went off the rails. Replicability Index Blog, February 2017.

Uri Simonsohn, Leif D Nelson, and Joseph P Simmons. P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534, 2014a.

Uri Simonsohn, Leif D Nelson, and Joseph P Simmons. p-curve and effect size:

correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9(6):666–681, 2014b.

Robbie CM van Aert, Jelte M Wicherts, and Marcel ALM van Assen. Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, 11 (5):713–729, 2016.

Marcel ALM van Assen, Robbie van Aert, and Jelte M Wicherts. Meta-analysis using effect size distributions of only statistically significant studies. *Psychological methods*, 20(3):293, 2015.