Does Psychotherapy Alleviate Depression?[*]

Johannes Haushofer[†][**], Daniel Mellow[‡], Pim Cuijpers[§]

May 9, 2019

Abstract

We analyze a database of psychotherapy RCTs using a set of techniques to test for and correct for publication bias, estimate power, and calculate post-experimental odds. In doing so

JEL codes:
Keywords: go here
Word count:

[*] 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.

[†] Department of Psychology, Woodrow Wilson School for Public and International Affairs, and Department of Economics, Princeton University; and The Busara Center for Behavioral Economics, Nairobi, Kenya. haushofer@princeton.edu.

[**] Corresponding author: 427 Peretsman Scully Hall, Princeton, NJ 08544, USA.

[‡] Department of Psychology, Princeton University; and The Busara Center for Behavioral Economics, Nairobi, Kenya. dmellow@princeton.edu

[§] Department of Clinical, Neuro and Development Psychology, VU Amsterdam, The Netherlands

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., 2013a], 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.

We follow a large body of recent work investigating the replicability of phenomena across the human subjects experimental sciences, notably social psychology and clinical medicine. In many cases, psychological effects with large bodies of literature supporting their existing have failed to replicate[Klein et al., 2014] . This is thought to be because journal editors and peer reviewers are more likely to deem an experimental paper worthy if the effect is statistically significant. It is hypothesized that researchers respond to this incentive structure by either abadoning projects that did not produce significant results [Simonsohn et al., 2014a], or manipulating the data until they find a test which allows them to reject the null hypothesis [Simmons et al., 2011]. In either situation, the experimental results in the published record will not consistute a representative sample of effect size estimates, but be biased upwards in absolute value.

In response to exposure over the scale of this problem, the academic community has been galvanized to improve research practices. However, these efforts cannot raise the evidential value of previously published studies. Recently a large literature has emerged of researchers attempting to solve the publication bias problem ex post by developing models which can, under some stylized conditions, recover the true effect size distribution from an empirical distribution that is possibly censored by a selection process representing publication bias or research questionable research practices (QRPs). Section 3 briefly summarizes this literature and selects models for implementation in the psychotherapy literature.

We find the psychotherapy literature is generally underpowered. Despite this, the publication bias-corrected estimates of effect size generally cluster around an effect size of .58 standard deviations, only slightly smaller than the 'naive' estimate of .72. Therefore cognitive-behavioral therapy for major depressive disorder avoids the fate of some other psychological phenomena when exposed to scrutiny. However, the results still indicate that the effectiveness of psychotherapy is overstated by the present literature.

The present study is most similar to other empirical applications of publication bias correction models [Simonsohn et al., 2014a,b, van Assen et al., 2015] and previous meta-analyses of the psychotherapy literatures using the same sample [Barth et al., 2013b, Cuijpers et al., 2010]. Of the latter type, two recent studies are most pertinent to the current paper.

Flint et al. [2015] demonstrate that the psychotherapy literature contains more significant results than would be expected given the (post hoc) power of the studies, though the main test is only marginally significant. We do not replicate this result, possibly because we have a larger sample of studies that may be subject to less publication bias.

Driessen et al. [2015] provide evidence for very strong publication bias in the psychotherapy literature among NIH-funded trials. They find huge differences between the average effect among published and unpublished studies, providing a direct test for the extent of publication bias in the psychotherapy literature. We cannot speak to the veracity of their main finding about publication bias, but when we apply meta-analytic methods for correcting for publication bias we observe a decrease in effect size that is similar in magnitude. However, our meta-analytic effect sizes are somewhat larger, likely owing to the difference sample selection.

This paper differs from all previous efforts in three ways. First, to our knowledge no meta-analysis of the psychotherapy literature attempts to correct for publication bias using the most modern techniques. Second, we present three of the best tools for dealing with study selection side by side. The fact that these fundamentally different approaches generally produce a similar estimate allows us to make a more definitive statement about the true strength of psychotherapy as a treatment for depression. Lastly, we augment these frequentist approaches with a Bayesian framework in which we calculate post hoc rejection ratios — the probability that a prior distribution is false over the probability that is is true – under a number of different priors to provide further evidence that the effect of psychotherapy on depression is real and large in size.

DM: here is some more interesting literature:

Cuijpers et al. [2019] analyze the same database and find that only 23% of studies are at low or no risk of bias according to the Cochrane guidelines. While the overall meta-analytic effect size is d=0.7, they find that the effect among the low-bias studies is only 0.31, much lower than previous estimates.

Munder et al. [2018] retort by saying that Waitlist Control studies should not be excluded, and that Cuijpers et al. [2019] do not apply the bias guidelines properly.

Renkewitz and Keiner [2018] compare several of the estimators we consider in a simulation framework. They compare p-uniform methods with traditional methods and two funnel plot-based methods we do not consider, and come to very similar results

2. Data

We obtain our sample from a database of randomized controlled trials comparing psychotherapy to a control condition, initially compiled by Cuijpers et al. [2008] and recently updated as of January 1, 2017. The database was developed by means of several methods. First, the aurthors conducted a comprehensive literature search (from 1966 to May 2007) in which they examined 5,178 abstracts in the following databases: Pubmed (1,224 abstracts), Psycinfo (1,736), Embase (1,911) and the Cochrane Central Register of Controlled Trials (2,056). Abstracts were identified by combining terms indicative of psychological treatment (psychotherapy, psychological treatment, cognitive therapy, behavior therapy, interpersonal therapy, reminiscence, life review) and depression (both MeSH-terms and text words).Cuijpers et al. [2008] also compiled the primary studies from 22 meta-analyses of psychological treatment of depression.

RCTs in the database are identified by type of thereapy devliered, the format of delivery, control condition, country of study, target population and the number of participants per group, among other charactersitics. In the case of studies with more than two treatment arms, each comparison is listed separately. Where possible, estimates have been converted to an effect size, resulting in a total of 329 effects from 253 studies included in this analysis. More detailed information can found at www.evidencebasedpsychotherapies.org.

3. Methods

Our analysis has been pre-specified with pre-analysis plan in the AEA RCT Registry, ID AEARCTR-0002574. We implement eight estimators in total, drawn from three classes of models: p-curve/p-uniform, parametric selection models, and the traditional methods that assume no publication bias.

Traditional mixed-level models [Borenstein et al., 2010]:

- Fixed-effects framework
- Random-effects framework

P-uniform class estimators:

- Kolmogorov-Smirnov approach (KSmirnov;Simonsohn et al. 2014b)
- Fisher statistic approach with $F = \sum_{i=1}^{N} - \ln(p_i)$ ($Ln(p)$;van Assen et al. 2015)
- Fisher statistic approach with $F = \sum_{i=1}^{N} - \ln(1 - p_i)$ ($Ln(1 - p)$;van Assen et al. 2015)
- Irwin-Hall appraoch (Irwin-Hall; van Aert et al. 2016)

Structural models of publication bias:

- Three parameter selection model [3PSM; Carter et al., 2017, McShane et al., 2016]
- Andrews-Kasy model[A-K; Andrews and Kasy, 2017].

3.1 Traditional Meta-analysis

For comparative purposes we implement the classic fixed and random effects estimators of meta-analytic effect sizes. The fixed-effect estimate is a weighted mean of effect sizes, using inverse-variance weights. The random-effect method allows for variation in the "true" effect size that forms the mean of the sampling distribution and therefore uses a weighting formula that incorporates the estimated variance of the latent distribution that generates the effect size for each study (which the sampling distribution for the estimate from that study is centered on). See Borenstein et al. [2010] for a more detailed introduction to traditional methods of meta-analysis.

The first wave of econometric methodologies, such as "Trim and Fill" [Duval and Tweedie, 2000] and the Precision Effect Test (PET) or Precision Effect

Test - Precision Effect Estimate with Standard Error (PET-PEESE) [Stanley and Doucouliagos, 2014], to correct for publication bias involved looking for ways in which the power of a study was related to the estimated effect. The core of this idea is that if noisier studies produce larger effect sizes on average, publication bias or other dark forces must be censoring the small or negative effects. These methodologies are intuitively compelling, but we avoid them for two reasons. One is that simulation evidence indicates that neither of the two most popular methods detailed below consistently recover the true mean of a latent distribution from a censored sample [van Assen et al., 2015, Carter et al., 2017]. Secondly, the models require an unrealistic assumption that the sample size of a study is unrelated to the size of the true effect studied, which is tantamount to declaring that all pre-experimental power calculations are meaningless.[1]

3.2 P-curve and P-uniform

The p-uniform approach [van Assen et al., 2015] is based on the fact that *if the null hypothesis is true* the p-values of hypothesis tests follow a standard 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, creating the conditions for a moment-matching estimator. The difference in the estimators in the classes arise solely from how to best measure the “closeness” of an empirical and ideal distribution in this context. The existing literature proposes 4 possible methods to do so:

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.

[1] See http://datacolada.org/58 and http://datacolada.org/59 for a fuller explanation of this critique.

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 the logtharmic function is highly sensitive to small differences close to 0, the authors recommend another estimator based on the *complements* of the p-values. We calculate both for completeness, but confirm that the latter does indeed produce a better estimator.

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. This remains the authors preferred approach.

We implement all four p-uniform methods using the `puniform` R package provided by Robbie C.M. van Aert.

Inference

Along with producing an effect size estimate that corrects for publication bias, the p-uniform approach also allows for straightforward extensions for inference. Confidence intervals are calculated by moment-matching: the lower bound of the two-sided $1 - \alpha$ confidence interval is the effect size $d_{\frac{\alpha}{2}}$ that produces a test statistic (Fisher or sum of p-values) which is at the $\frac{\alpha}{2}$ quantile of the null distribution, and vice versa for the upper bound.

Note that there is no clear analoge for the Kolmogorov-Smirnov method, so we do not calculate any confidence intervals for this appraoch. However, the Kolmogorov-Smirnov statistic is useful for testing for the existence of publication bias. Under the p-uniform assumptions, the null hypothesis of no publication bias is equivalent to the hypothesis that the mean of the observed studies produces a uniform distribution of conditional p-values. This can then by tested using a standard implementation of the Kolmogorov-Smirnvo test.

Another important consideration for implementation of the p-uniform methods is the assumptions that each effect size estimate is from an independent sample. In the case where one experiment produces multiple treatment arm comparisons relative to the same control group, we randomly select one for inclusion in our sample for p-uniform estimates.

3.3 Parametric Selection Models

We consider two methods which attempt to explicitly model the publication bias process. The three-parameter selection model [3PSM; 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 a discontinuity at $p = 0.05$. If effect sizes are standardized, this model is characterized by three parameters:

1. The mean true effect size, μ.
2. The ratio of the probability of publication of nonsignificant studies to that of statistically significant studies, $\beta \in [0, 1]$.
3. A heterogeneity parameter, τ, equal to the standard deviation of the sampling distribution of underlying true effect sizes.

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 theoretically equivalent to the foundation of p-curve/p-uniform. The only difference is in identification strategy: p-uniform utilizes moment matching, whereas the 3PSM estimates parameters by maximum likelihood. Since maximum likelihood estimation is 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 implement the 3PSM using the custom R function available in the supplementary material of McShane et al. [2016].

Secondly, we 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 allow 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 implement the A-K estimator using a custom R package, available from the authors on request.

3.4 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 and Andrews-Kasy estimators will be lower bounds.

We 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. If the effect estimated from the empirical data is above the 97.5 percentile of the simulated distribution, we reject the null hypothesis.

For robustness, we conduct simulations under different eight different data generating processes. Each simulation environment is characterized by: (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," using an algorithmic framework for simulating these behaviors in the creation of meta-analytic samples created by Carter et al. [2017].

4. Results

Figure I shows our estimates of meta-analytic effect size, and the implied middle 50% of post hoc power implied by each. Panel A illustrates that the effect size estimates generally converge to a narrow band around $d \approx 0.6$. In particular, it is noteworthy that the fixed-effect estimate ist not meaningfully different from the 3PSM and Andrews-Kasy estimates, and below three of the four p-curve/p-uniform estimates. This result is in contrast to Driessen et al. [2015]. We assume the difference is due to the selection of sample, as we do not have direct access to unpublished experiments.

There are two exceptions to the general pattern of estimates bunched close together. First, the Fisher statistic estimator of the p-uniform model, labelled $Ln(p)$, is implausibly high at $\hat{d}_{Ln(p)} = 1.19$. This is one instance of a consistent pattern discussed in section 4.1 below. Second, the random-effect estimator, implemented using the maximum-likelihood procedure, is meaningful greater than most of the other estimates at $\hat{d}_{RE} = 0.72$, though it is not very precise and therefore statistically indistinguishable from the Irwin-Hall and Kolmogorov-Smirnov estimates. That the random-effect estimator is meaninfully higher than the fixed-effect is indicative of a large amount of between-study variation and less precise studies producing larger effects. Indeed, the Q-test of between-studies heterogeneity overwhelming rejects the null hypothesis that all studies have the same underlying effect ($Q_{328} = 1218, p < 0.0001$). This is to be anticipated, given the studies in our sample include a wide variety of interventions, target populations, study sites, control conditions and outcome measures.

Since we presume that the effect of psychotherapy varies depending on the type of intervention, we conduct the same analysis above for the different types of psychotherapy in the sample. Figure II displays the estimates for effect size separately for each type of psychotherapy. Note that the subgroup analyses still pool by format, control condition and study location within type of psychotherapy. As the majority of the studies in our sample are tests of cognitive-behavioral therapy, the CBT results are qualitatively similar to the pooled results, while many of the other types suffer the imprecise estimates due to a small sample.

Behavioral-activation therapy (BAT) is supported by five of six estimators despite a small sample size, as are interpersonal (IPT) and problem-solving therapies (PST). Supportive therapy (SUP) appears to be precisely estimated to be effectively, but less so than the other types in magnitude. Psychodynamic therapy (DYN; Panel F) alone appears to be ambiguously significant: four of the six bias-corrected estimates cannot reject the null. Lastly, mindfulness-based cognitive therapy (MBCT) is represented by only seven studies in our sample; despite the resulting imprecision, the overall effect of MBCT appears to be large.

We also group together 57 studies that do not belong to any of the other groups. Many of these studies present a novel type of psychotherapy that has not gained currency in the wider clinical or academic communities. For example, Puckering et al. [2010] test the effectiveness of a particular institution's postnatal intervention, called "Mellow Babies", while Carlbring et al. [2013] propose a mix of behavioral-activation therapy and acceptance and commitment therapy to be conducted online. The effect sizes across this group of studies are qualitatively similar to those of CBT.

4.1 How good are our meta-analytic methods?

Figure II illustrates how the empirical estimates of effect size compare to their simulated null distributions. We simulate eight different distributions. In each of these, the true effect size of the intervention is zero, but the literature is subject to different degrees of distortion along two dimensions: heterogeneity of underlying true effect size and the degree to which researchers engage in some reportedly common questionable research practices (QRPs), including publication bias. The number of studies and their sample sizes are taken from the empirical database, but all other data is simulated. Note that Panel A includes confidence intervals for the empirical estimations, as it is under the conditions of that simulation that the estimated standard errors are consistent.

We find that, though under some conditions all of the estimators can be biased and/or very noisy, it is only under the most extreme distortions of the data generating process that the empirical estimates are within the bounds of the simulated distribution under no effect. In other words, we would have to be deeply cynical about the state of the publication process and researcher behavior for our empirical estimates to be generated simply by random chance. In particular, only when both heterogeneity is high and publication bias omnipresent that the simulated distributions of the p-curve/p-uniform estimators approach their psychotherapy means. Following this logic, we can state with a high degree of certainty that clinical psychotherapies are still valid in the face of scrutiny concerning publication bias.

Figure III illustrates the comparative performance of each of our methods under a variety of data generating processes and true effect sizes. We consistantly observe that the Fisher statistic estimator of the p-curve model (labeled $Ln(p)$) is upwardly biased when heterogeneity of true effect size is present. This problem is common across all the p-curve/p-uniform estimators, but worst for the $Ln(p)$ method, including when there really is nonzero mean true effect. Since it is almost certainly true that our sample is characterized by a high degree of heterogeneity in effects ($\tau^2 = 0.21, I^2 = 78.7\%$), we must view the $Ln(p)$ estimator, and all the p-curve/p-uniform generally, as upper bounds. Why the $Ln(p)$ estimator performs much worse than the other p-curve/p-uniform methods is not clear, although van Assen et al. [2015] argue that because the logarithmic function is highly sensitive to small changes close to zero the estimator is likely to be noisy and sensitive to rounding errors. Our findings largely confirm this interpretation. We continue to present the $Ln(p)$ estimator for completeness and compliance with our pre-analysis plan.

Estimator Bias as a Function of Heterogeneity and Questionable Research Practices

To further explore the bias in estimators visible in Figure III, we run regressions of the form

$$\hat{d}_i - \delta_i = \alpha_0 + \alpha_1 \delta_i + \alpha_2 \tau_i + \sum_{j=1}^{J=2} \gamma_j QRP_{ji} + \alpha_3 \tau_i \cdot QRP_i$$

Where $\hat{d}_i$ is the estimated effect size for iteration i; $\delta_i \in \{0, .2, .4, \ldots, 1\}$ is the mean true effect size; $\tau_i \in \{0, .2, .5\}$ is the standard deviation of the sampling distribution of true effect sizes, such that in each iteration the true effect $d_i \sim N(\delta, \tau^2)$; $QRP \in \{none, med, high\}$ is the level of questionable research practices in the simulation environment. Table I displays regression results. Each column represented simulation results for a different method. The dependent variable is the difference between the estimated and true effect size.

Note that the intercept term, α_0 is equal to the average effect estimate under zero true effect and ideal conditions of no heterogeneity and no QRPs; therefore α_0 should be equal to zero if the estimator is unbiased. However, all methods display some bias even in this simple situation. The traditional methods display a slight upward bias at zero true effect while the rest slightly underestimate the true effect. All converge towards to the true estimate as δ increases, however. The case of the $Ln(1-p)$ estimator is especially intriguing: in environments of high heterogeneity, the slope of the estimates with respect to δ is less than one, meaning that the method displays upward bias when the true effect size is close to zero and downward bias when the true effect size is high. This dynamic is clearly visible in the rightmost panels of Figure III.

As might be expected, the traditional meta-analytic methods are susceptible to QRPs, while the publication-bias-robust estimators are not meaningfully affected by QRPs alone. The effects of QRPs on p-uniform methods are large in magnitude but not statistically significant. A high QRP environment does have statistically significant effects on 3PSM and A-K, but the magnitudes are small.

Heterogeneity unsurprisingly causes massive upward bias in p-uniform estimators, as reported by methods authors, to the point that these methods are essentially invalidated as τ^2 approaches .25. The rest of the methods are generally robust to heterogeneity. This is particularly interesting in the case of fixed-effects, which assumes no underlying heterogeneity but is not systematically biased by the presence of even high heterogeneity, probably because the heterogeneity in these simulations is symmetric.

The /emphinteraction of heterogeneity and QRPs produces strong downward bias in both A-K and 3PSM, as well as upward bias in FE and RE. The additon of questionable research practices decreases the effect of large heterogeneity on p-uniform methods. This is evident in Figure III, where p-uniform estimators are less biased on average in Panel I than Panel C.

We broadly summarize the performance of our methods as follows: p-uniform methods are robust to QRPs but highly vulnerable to heterogeneity, while traditional methods handle heterogeneity well but become upwardly biased in the face of QRPs. Selection models are relatively unaffected by heterogeneity and QRPs independently but break down when both are present. Since no estimators perform well in this worst case scenario, the simulations large support the theoretical argument that structural models of publication bias are the best method of correcting for them.

4.2 Bias Corrected Estimates

Simulation results and precise knowledge of estimator bias under a range of conditions provide an opportunity to augment our empirical estimates of psychotherapy effectiveness by correcting for the bias of the estimator. We conduct additional simulations with (i) empirically estimated τ, the heterogeneity parameter, for each subtype of psychotherapy; (ii) a “medium” level of QRPs; (iii) 100 studies per iteration. We proceed to back out “corrected” effect estimates by finding the true δ, to the nearest 0.05, that produces the mean simulated estimate $\bar{d}$ closest to the empirical estimate for each combination of psychotherapy and method.

Figure IV presents these results. In all 9 samples, FE and RE estimates are corrected downwards. In Panel A, including all 329 studies, the corrected FE estimate is the lowest at 0.35. The selection models A-K and 3PSM, on the contrary, are corrected upwards in all samples, likely due to the high levels of heterogeneity even within psychotherapy types. P-uniform methods are also generally corrected down, except for $Ln(1-p)$, which is sometimes corrected up, albeit by small amounts.

To arrive at a final estimate for the effect of psychotherapy on depression, we selectthe least-biased estimator in the full sample $Ln(1-p)$ and the corresponding corrected effect size: $d = 0.6$.

5. Discussion

Despite growing concerns about the evidential value of the psychotherapy literature, applying the best methods to correct for publication bias does not change drastically alter the estimates. Our results do not substantiate the claim that the apparent effectiveness of psychotherapy for depression is an artefact of publication bias or researcher shenanigans.

In the process of applying the new wave of methodologies to correct for publication bias, we find that these methods all have weakness. Simulation data indicate both p-uniform methods and structural selection models reliably reject the null, though in opposite directions, when there is no true effect, even under ideal data generating processes. As we add the complications of heterogeneity of true effect size, publication bias and p-hacking to the simulation environment, our methods all suffer from increased bias. Specifically, traditional meta-analytic methods are biased upwards by questionable research practices but are unaffected by heterogeneity in effect size; p-uniform approaches generally handle QRPs well but become highly biased in the presence of heterogeneity; selection models handle these two forces well independently, but display strong downward bias when both are present.

However, for the range of heterogeneity, true effect size and questionable research practices that are plasuible for this literature, the empirical effects are well outside the range of simulated null distributions (that is, simulations with no true effect size), implying that we can confidently reject a null hypothesis of no tue effect.

Still, the issue of bias in our estimators is disconcerting. We adopt a strategy of explicitly adjusting for this bias. Setting simulation parameters to their empirical or most plausible values, we back out the true effect size that produces estimates closest to what we get from the psychotherpy literature.

6. Conclusion

Goes here.

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 2013a. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001454. URL `http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001454`.

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, 2013b.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, 1(2):97–111, 2010.

Per Carlbring, Malin Hägglund, Anne Luthström, Mats Dahlin, Åsa Kadowaki, Kristofer Vernmark, and Gerhard Andersson. Internet-based behavioral activation and acceptance-based treatment for depression: a randomized controlled trial. *Journal of Affective Disorders*, 148(2-3):331–337, 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.

P. Cuijpers, E. Karyotaki, M. Reijnders, and D. D. Ebert. Was eysenck right after all? a reassessment of the effects of psychotherapy for adult depression. *Epidemiology and Psychiatric Sciences*, 28(1):21â30, 2019. doi: 10.1017/S2045796018000057.

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`.

Pim Cuijpers, Filip Smit, Ernst Bohlmeijer, Steven D. Hollon, and Gerhard Andersson. Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: meta-analytic study of publication bias. *The British Journal of Psychiatry*, 196(3):173–178, March 2010. ISSN 0007-1250, 1472-1465. doi: 10.1192/bjp.bp.109.066001. URL `http://bjp.rcpsych.org/content/196/3/173`.

Ellen Driessen, Steven D Hollon, Claudi LH Bockting, Pim Cuijpers, and Erick H Turner. Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? a systematic review and meta-analysis of us national institutes of health-funded trials. *PLoS One*, 10(9):e0137864, 2015.

Sue Duval and Richard Tweedie. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2):455–463, 2000.

J Flint, P Cuijpers, J Horder, SL Koole, and MR Munafò. Is there an excess of significant findings in published studies of psychotherapy for depression? *Psychological medicine*, 45(2):439–446, 2015.

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.

Richard Klein, Kate Ratliff, Michelangelo Vianello, Reginald Adams Jr, Stěpán Bahník, Michael Bernstein, Konrad Bocian, Mark Brandt, Beach Brooks, Claudia Brumbaugh, et al. Data from investigating variation in replicability: A "many labs" replication project. *Journal of Open Psychology Data*, 2(1), 2014.

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.

Thomas Munder, Christoph Flückiger, Falk Leichsenring, Allan Anthony Abbass, Mark J. Hilsenroth, Patrick J Luyten, Sven Rabung, Christiane Steinert, and Bruce E. Wampold. Is psychotherapy effective? a re-analysis of treatments for depression. *Epidemiology and psychiatric sciences*, pages 1–7, 2018.

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

Christine Puckering, Emily McIntosh, Anne Hickey, and Janice Longford. Mellow babies: a group intervention for infants and mothers experiencing postnatal depression. *Counselling Psychology Review*, 25(1):28–40, 2010.

Frank Renkewitz and Melanie Keiner. How to detect publication bias in psychological research? a comparative evaluation of six statistical methods. *PsyArXiv Preprints*, 2018.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

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.

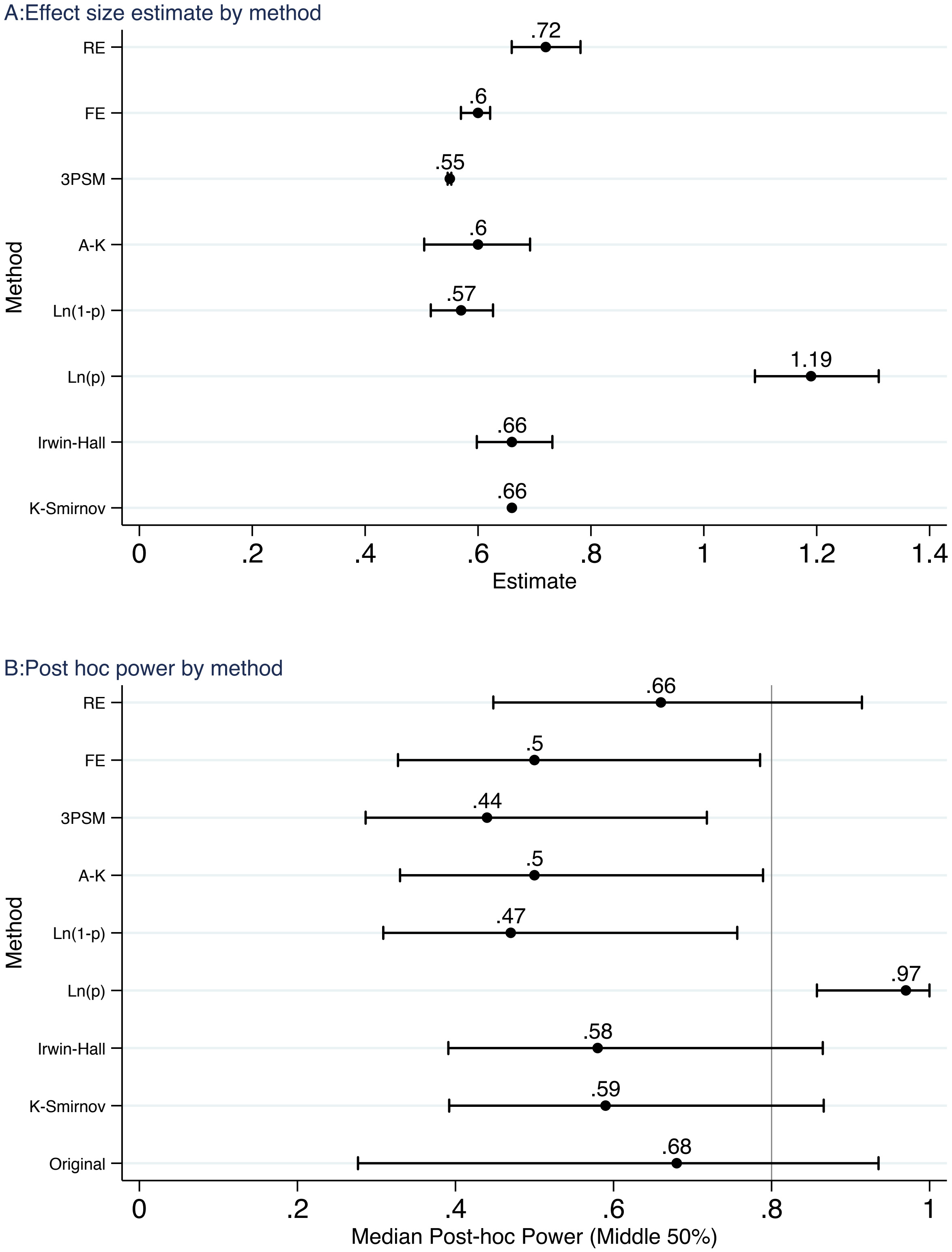
Uri Simonsohn, Leif D Nelson, and Joseph P Simmons. p-curve and effect size: correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9(6):666–681, 2014b.

TD Stanley and Hristos Doucouliagos. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1):60–78, 2014.

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.

Figure I: Estimated Meta Analytic Effect Size and Post Hoc Power by Method

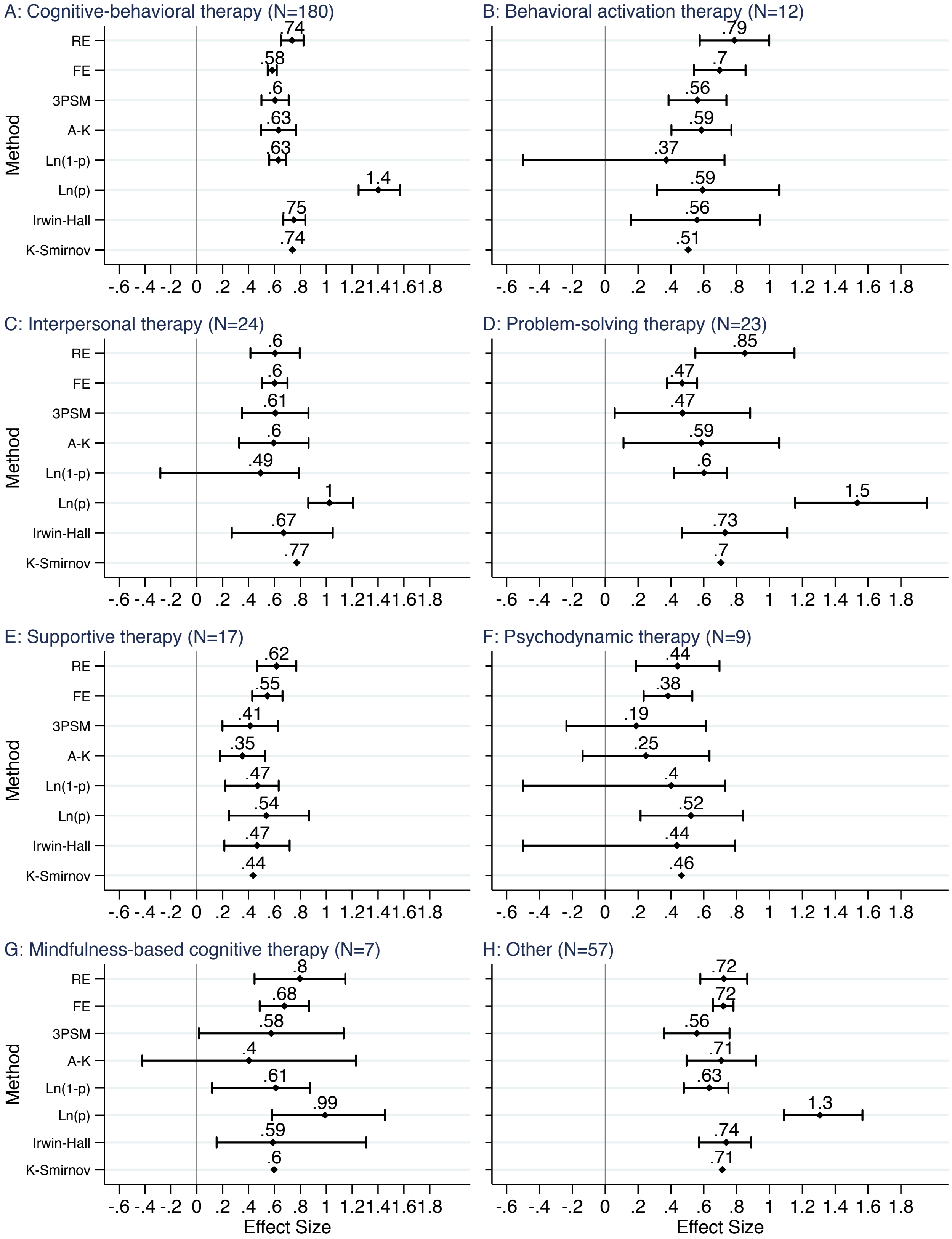


A:Effect size estimate by method

B:Post hoc power by method

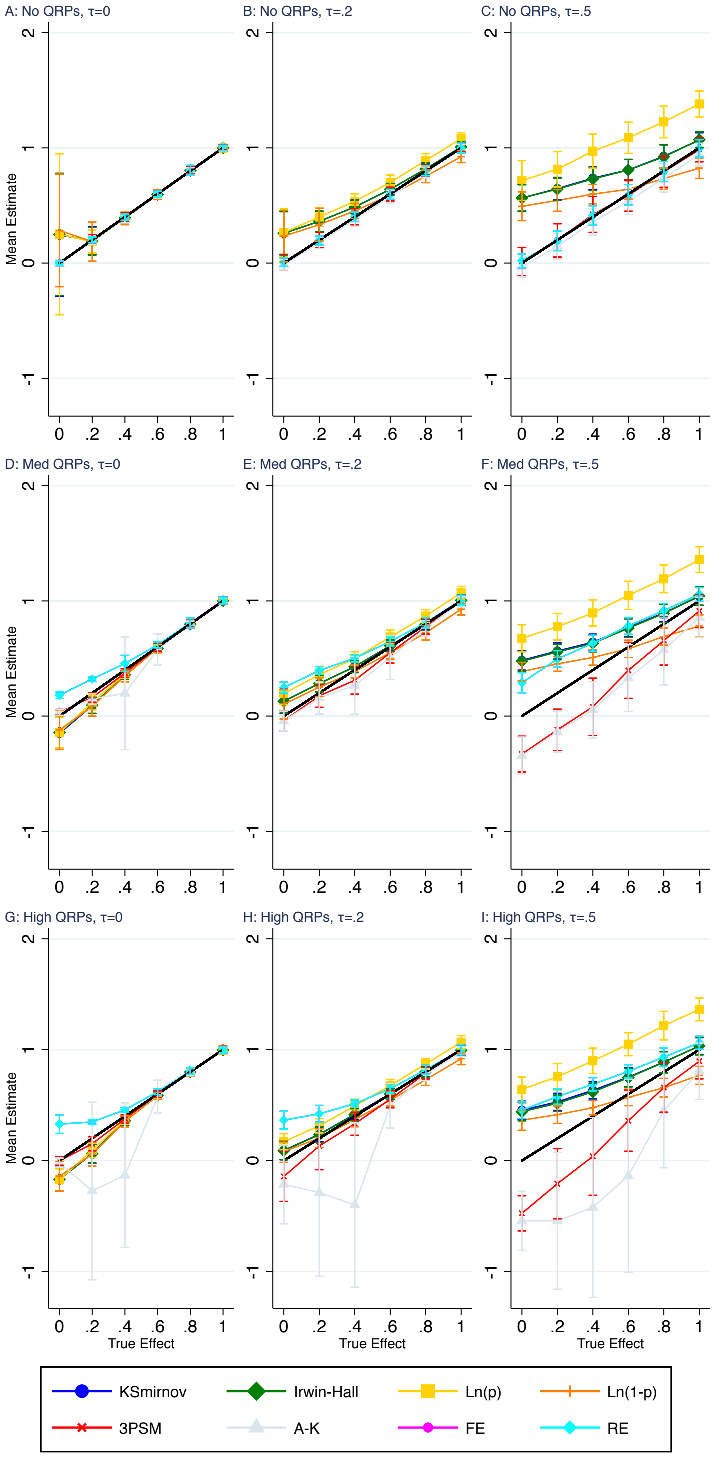
Notes: In Panel B, "original" refers to the post hoc power conditional on the estimated effect size for each study.

Figure II: Estimated Meta Analytic Effect Size by Method and Type of Therapy



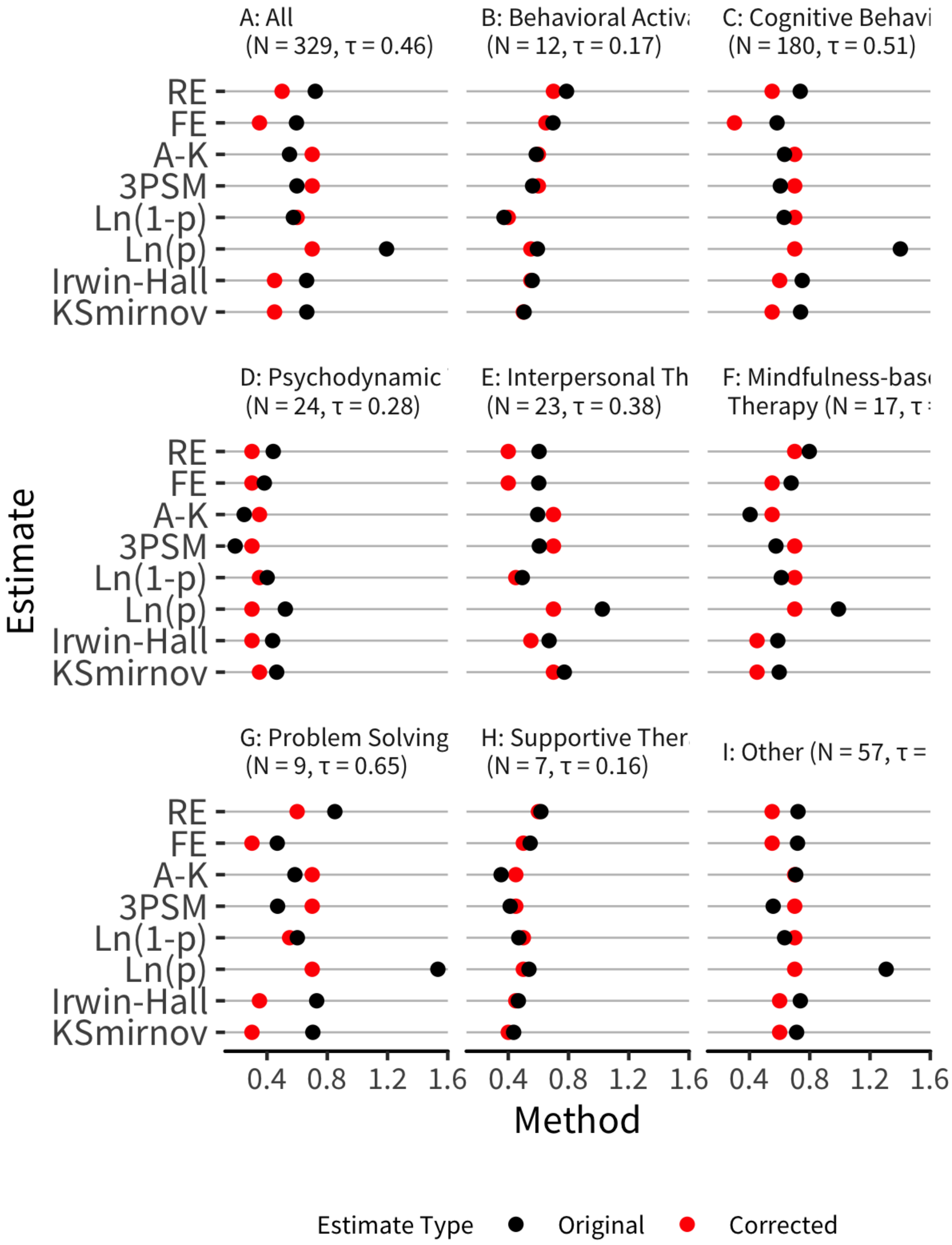
Notes: Kolmogorov-Smirnov method has no mechanism for inference. Number of studies in parenthses.

Figure III: Simulated Accuracy of Estimators



Notes: Points represent the mean of estimated effects from 100 iterations at each parameter combination, with 100 studies simulated per iteration. Error bars represent one standard deviation above and below the mean.

Figure IV: Simulated Accuracy of Estimators



Notes: Correction factors are estimated using a simulation framework with a ”medium” degree of questionable research practices, 100 studies per iteration and the empirical level of heterogeneity among the subtypes of psychotherapy.

Table I: Estimator Bias, as Function of Heterogeneity and QRPss

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	KSmirnov	Irwin-Hall	Ln(p)	Ln(1-p)	3PSM	A-K	FE	RE
True Effect(δ)	0.29	0.29	0.55	0.05	0.11***	0.10***	−0.19***	−0.19***
	(0.29)	(0.29)	(0.41)	(0.19)	(0.00)	(0.01)	(0.00)	(0.00)
Heterogeneity (τ)	2.35**	2.34**	3.65**	1.43**	−0.00	−0.07*	−0.00	−0.00
	(0.83)	(0.83)	(1.20)	(0.55)	(0.01)	(0.03)	(0.00)	(0.00)
Med. QRP	0.67	0.67	1.01	0.43	0.02***	0.01	0.06***	0.06***
	(0.36)	(0.36)	(0.53)	(0.24)	(0.00)	(0.01)	(0.00)	(0.00)
High QRP	0.66	0.66	1.00	0.42	0.01***	−0.05**	0.08***	0.08***
	(0.36)	(0.36)	(0.53)	(0.24)	(0.00)	(0.01)	(0.00)	(0.00)
$\tau\, X\, Med.QRP$	−1.76	−1.77	−2.53	−1.19	−0.47***	−0.46***	0.28***	0.28***
	(1.17)	(1.17)	(1.69)	(0.77)	(0.01)	(0.05)	(0.01)	(0.01)
$\tau\, X\, High.QRP$	−1.78	−1.80	−2.54	−1.22	−0.71***	−1.20***	0.34***	0.34***
	(1.17)	(1.17)	(1.69)	(0.77)	(0.01)	(0.05)	(0.01)	(0.01)
Constant	−0.88**	−0.88**	−1.36**	−0.51**	−0.05***	−0.05***	0.10***	0.10***
	(0.30)	(0.30)	(0.43)	(0.20)	(0.00)	(0.01)	(0.00)	(0.00)
Observations	16168	16168	16168	16168	16200	16200	16189	16189

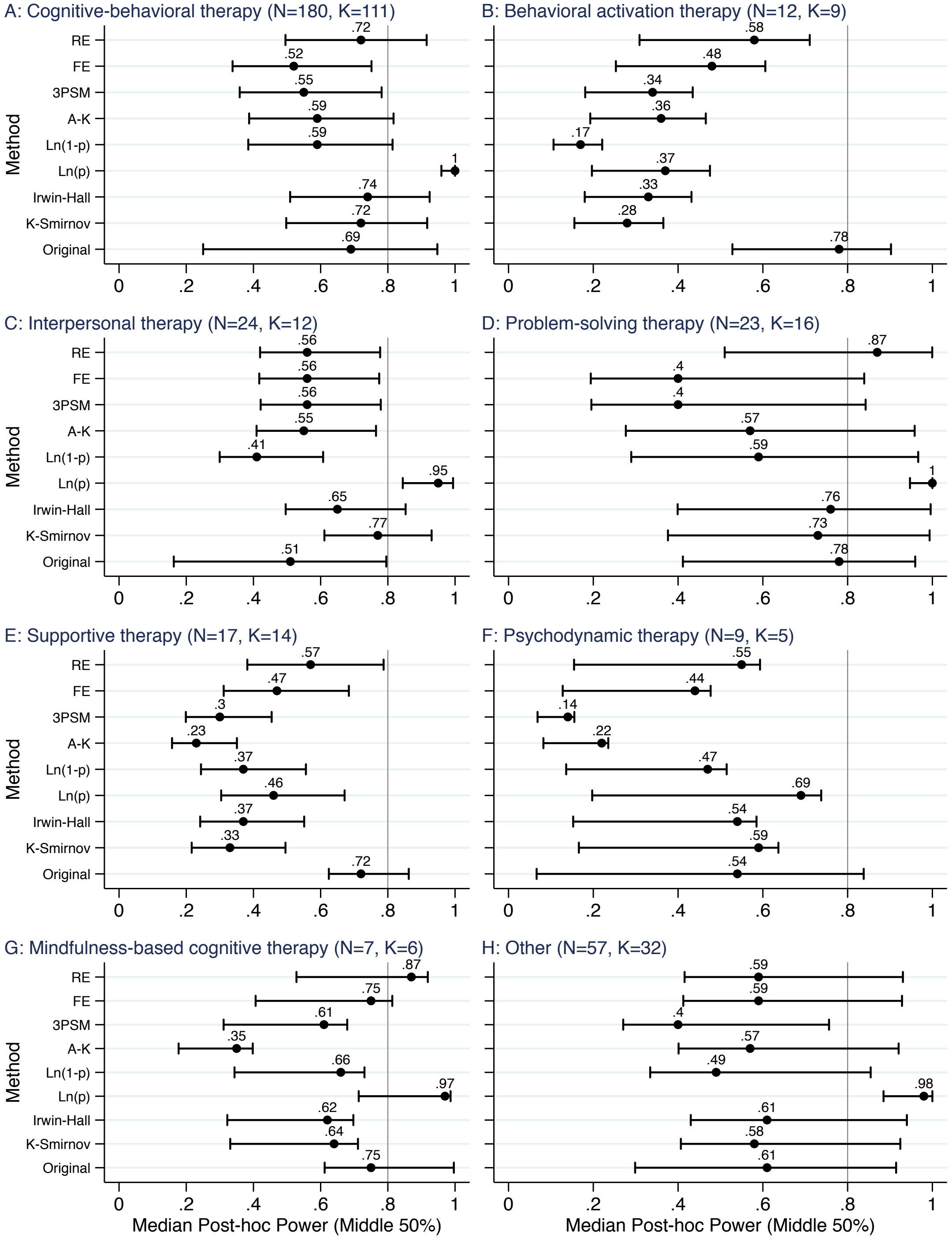
Notes: Dependent variable is equal to the difference between estimated effect and the true mean effect in each iteration of the simulation.

Appendix

A. Proofs

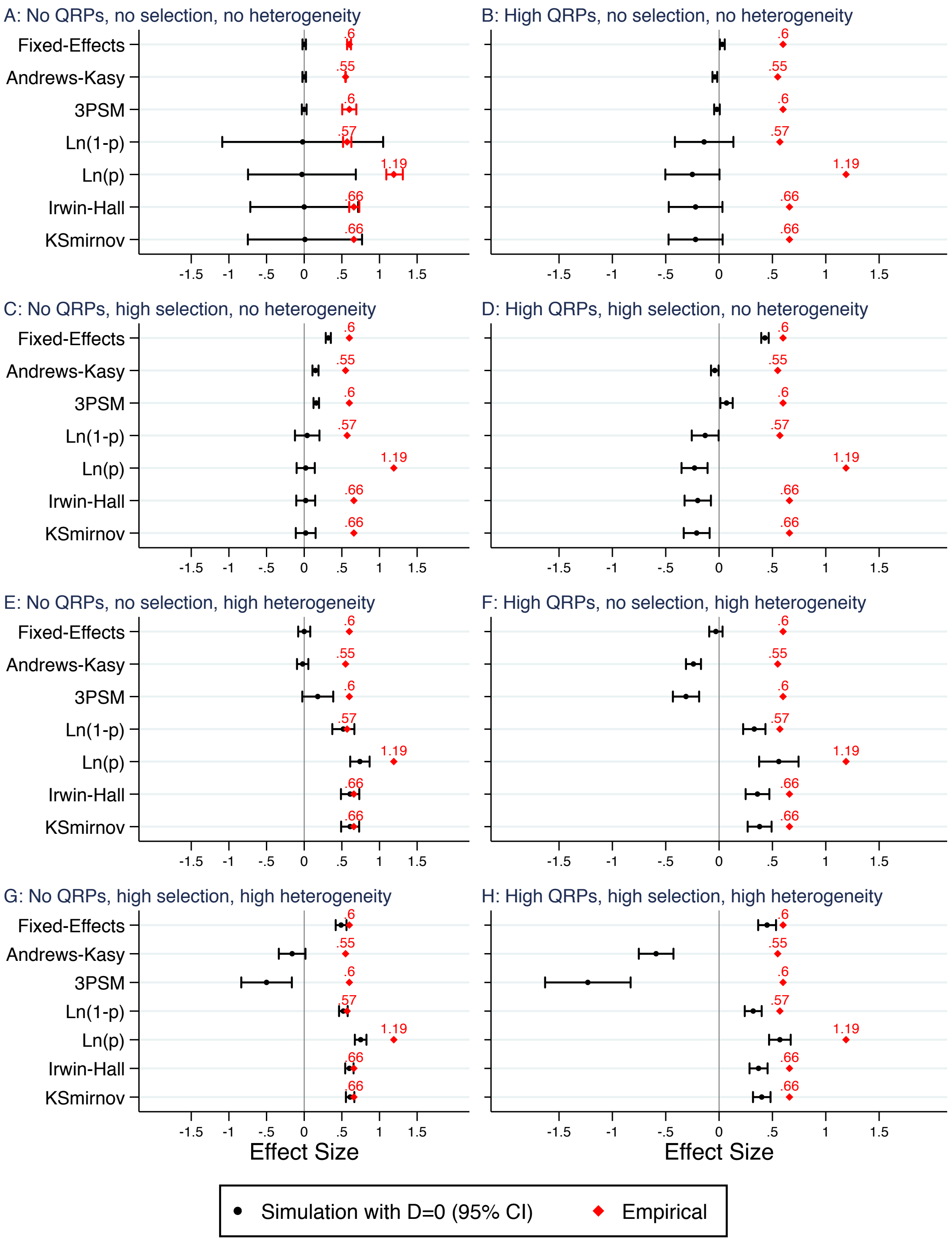
B. Additional Tables and Figures

Figure I: Post Hoc Power by Method and Type of Therapy



Notes: Line at 0.8 indicates the level of power conventionally deemed adequate. Upper and lower bounds represent the .75 and .25 quartile of the post-hoc power distribution, respectively.

Figure II: Distribution of Effect Size Estimates, by Method and Simulation Environment



A: No QRPs, no selection, no heterogeneity

B: High QRPs, no selection, no heterogeneity

C: No QRPs, high selection, no heterogeneity

D: High QRPs, high selection, no heterogeneity

E: No QRPs, no selection, high heterogeneity

F: High QRPs, no selection, high heterogeneity

G: No QRPs, high selection, high heterogeneity

H: High QRPs, high selection, high heterogeneity

Notes: Each panel represents a different simulated environment. Each environment is simulated 1000 times. Black lines denote the middle 95