Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A case study on applying p-curve analysis to clinical trial data

Gustavo Andres Marquez Flores
Vrije Universiteit Amsterdam
Luca Clivio
Amsterdam University Medical Center, University of Amsterdam
Alexios-Fotios A. Mentis
BGI-Shenzhen
Mariska M. G. Leeflang
m.m.leeflang@amsterdamumc.nl

Amsterdam University Medical Center, University of Amsterdam

Research Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11220384/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Title: A case study on applying p-curve analysis to clinical trial data.

Authors: Gustavo Andres Marquez Flores[1,2], Luca Clivio[2], Alexios-Fotios A. Mentis [3], Mariska M.G. Leeflang [2]

Author affiliations:

1: Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

2: Department of Epidemiology and Data Science, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

3: BGI-Shenzhen, Shenzhen, Guangdong, China

Corresponding author: Mariska M. G. Leeflang

Postal address: Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands

Email address: m.m.leeflang@amsterdamumc.nl

Fax number: N/A

Phone number: +31(0)620061752

Abstract

Background: Publication bias may occur when smaller studies reporting statistical significance ($p < 0.05$) are more likely to be published than smaller studies that do not report statistical significance. One way to detect publication bias may therefore be to detect a skewed distribution of reported p-values in a set of published articles.

Aim: To showcase the complexity and uncertainty of p-curve analysis as a method to detect publication bias.

Methods: We re-analyzed a meta-analysis summarizing randomized controlled trials of treatment for depression, using p-curve analysis. One significant p-value per study was selected for the main analysis, with alternative significant p-values allocated to a robustness analysis, which we called *"robust p-curve"*. We used the *"p-curve web application"* (https://www.p-curve.com/app4/).

Results: Our dataset included 36 test statistics from 29 of 41 (70.73%) studies in the meta-analysis. The corresponding p-curve was neither significantly right-skewed (z = -1.03, p = .15) nor significantly flatter than a curve powered at 33% (z = -1.02, p = .15); it was, therefore, inconclusive. The robust p-curve was similar, though its right-skewness test was not statistically significant either (z = -1.61, p = .05; flatness z = -.55, p = .29). Both p-curves suggested that the average power of the included studies was low [Estimate: 14%, 90%CI (5%, 48%); Estimate: 22%, 90%CI (5%, 57%)].

Conclusions: Based on p-curve analyses, we could not rule out or rule in publication bias. One of the reasons may be that the sample size of the studies included was too small. P-curve analysis complements, rather than substitutes for, other methods for investigating publication bias. Our study

45 highlights the assumptions and uncertainty that researchers must deal with and reflect on when
46 analyzing potential publication bias.
47

Background

Publication bias occurs because small-in-size, low-powered studies with large or statistically significant effects that usually support the tested hypothesis are more likely to be submitted and accepted for publication than small-in-size, low-powered studies with small effects or effects that are not statistically significant (1). We are also considering publication bias when researchers only publish subsets of results that support a hypothesis, leaving in the file drawer those that do not (15). A p-value that is considered statistically significant ($p < .05$) seems to render results more attractive for publication to authors, reviewers, and editors (2, 3). Accordingly, the estimated frequency of results that supported the tested hypothesis increased by 22% between 1990 and 2007; of note, psychology/psychiatry and clinical medicine were two fields among those with the highest trends in question (4). In addition, publication bias is considered to be one of the largest threats to the validity of systematic reviews and meta-analyses (5) because it leads to overestimation of effect sizes (6).

Methods to evaluate the presence of potential publication bias often focus on the size of the effect. For instance, the *"trim-and-fill"* procedure (7), a method based on the funnel plot (8), which shows the results from the included studies in a meta-analysis, according to the variance of the study. It resembles a symmetric funnel around the true effect on the y-axis; hence, asymmetry should indicate publication bias. "Trim-and-fill" means that the asymmetric part of the plot is trimmed to estimate an "unbiased" center. Then, the plot is filled around the said center with the previously removed studies that were in the asymmetric part, together with imputed missing counterparts (7). Although publication bias methods based on the funnel plot are very popular (9), these may be focusing on testing between-study heterogeneity instead of publication bias (10).

In a recent meta-analysis, due to the shortcomings of funnel-plot based methods, the authors used two additional methods to investigate publication bias (11). First, they used *"limit meta-analysis"*, with which an additional parameter has to be included to adjust for the usually larger effects observed in small studies in an extended random-effects model meta-analysis (12). Second, the *"step function selection model"*, and, more concretely, the so-called *"three-parameter selection model"* was used.

Using this tool, researchers can adjust through a step function an effect size distribution for publication bias (13, 14); in other words, a relative likelihood is assigned to bins of p-values. To further illustrate this approach, it could be mentioned that significant p-values would have a higher selection probability to non-significant ones, and with that, an adjusted estimation can be made (14). When the original authors reported all three methods–i.e., "trim-and-fill", "limit meta-analysis", and "three-parameter selection model"–to provide comparisons between psychotherapies, two of them provided a small, but statistically significant effect, while the result of limit meta-analysis was non-significant.

In the literature, a fourth method to investigate publication bias has been presented, i.e., the p-curve approach. It focuses on statistical significance by assessing the distribution of a group of significant p-values (15). The shape of the distribution indicates whether the cluster of findings can be rejected as a likely consequence of reporting bias. A *"right-skewed"* p-curve suggests that the group of studies contains evidence of a real effect and thus that selective reporting or p-hacking can be ruled out. In contrast, a *"left-skewed"* distribution would indicate that it is not possible to reject selective reporting as the cause of significant findings. An advantage of the p-curve approach is the estimation of the average statistical power of the included studies with which one can examine whether the set of studies was too underpowered (or not) to detect negligible effect size (15, 16, 17, 18).

In the present study, we used p-curve to test for publication bias in the aforementioned meta-analysis by Cuijpers et al. 2024 (11) and, specifically, their results about comparing multiple treatments for depression. To this end, we here-in (a). analyzed the significant p-values of the clinical trials collected, (b). estimated the average statistical power of the studies showing significant results to determine whether they were too underpowered to detect negligible effect sizes, and (c). explored if publication bias can be *"ruled out"* as one of the major reasons for the retrieved findings. Our main aim has been to showcase the complexity and uncertainty of publication bias analysis more broadly using p-curve analysis.

2. Methods

2.1 Data sources

We harnessed data from the 41 randomized clinical trials included in the original meta-analysis by Cuijpers *et al.* (11). The authors selected trials published in English, except for one in German, from four different bibliographic databases (i.e., "Cochrane Library", "Embase", "PsycINFO", and "PubMed") between the 1st of January 1966, and the 1st of January 2023. (The search strings can be found on the project's website (www.metapsy.org and https://docs.metapsy.org/databases/)). The lead author provided us with the article's dataset on the 10th of June 2024.

2.2 Data compilation and extraction.

Because p-curve requires precise p-values, we collected the test statistics from which the reported p-values are derived (20). The following test statistics were gathered: Chi-squared test, *t*-test, *F*-test, Pearson's correlation test, and regression coefficients. We transformed the latter into *t*-values by dividing the regression coefficient by the standard error of the coefficient, again, as *t*-values are required by the tool we applied for the analysis. If researchers did not report the test statistics, data were retrieved from the report to recalculate them, e.g., means, sample size, standard error or others.

As p-curve analysis has been designed to analyze results under the assumption that statistical significance drives publication bias, only statistically significant results have been included (23). We extracted significant results for different measurements of depression, similarly to what the authors of the original meta-analysis performed (11). If a report contained two measurements with significant results, we selected only one p-value with a set of choices as seen in **Supplementary Table 1**.

Supplementary Table 1 contains *a)*. the article's stated hypothesis; *b)*. the design of the randomized controlled trial—i.e., interaction effect, linear trends, number of compared groups, and so on; *c)*. the statistical result used to test the hypothesis; *d)*. whether the main hypothesis involved testing the superiority of any psychotherapy over Non-Directive Supportive Therapy (NDST), which is the

treatment whose effectiveness was analyzed for depression compared to other psychotherapies in the original meta-analysis (11); *e)*. a quote from each article indicating where from the original reports the data was extracted from; *f)*. the test statistics for the main analysis; *g*). a quote from each article indicating from where in the original reports the data was extracted from to be in the robust analysis; *h)*. the test statistic for robust analysis; and, *i)*. additional notes that give our rationale for our selection of p-values. The first author (GAMF) filled in the p-curve-related Supplementary Table. Two authors (AFAM and MMGL) scrutinized 8 (20%) randomly selected studies as a validity and reliability assessment. Finally, the remaining author (LC) made a final check of all of the data and of the manuscript.

As it is not possible to analyze non-significant p-values using the p-curve method, test statistics that were incorrectly reported as significant were excluded. We used all our gathered data on p-values—i.e., non-significant, non-reported, reported p-value without test-statistics for plotting a final graph.

2.3 Analysis

We used the p-curve web application (https://www.p-curve.com/app4/) version 4.10, last updated in August of 2024. P-curve analysis estimates the probability of each p-value being observed under the assumption that it belongs to a *"no-effect"* distribution. The analysis consists of creating new p-values out of the extracted ones. As these new p-values get lower, their probability of belonging to a *"non-effect"* distribution gets lower as well (15, 17). Next, these new p-values are aggregated into a single, combined p-value, according to Stouffer(17). Then we evaluate whether the null hypothesis of all p-values belonging to a *"non-effect"* distribution—a uniform distribution—should be rejected in favor of its alternative that posits the existence of an effect—a *"right-skewed"* distribution. If the null hypothesis is rejected, we assume that selective reporting can be ruled out (15, 16).

As uniform p-curves may also indicate a noisy estimation due to, for example, too few p-values in the analysis (15), the obtained curve is also compared to one that represents the distribution of studies

with average power. This average power is 33%, based on relevant non-central distributions (15). If the curve of the extracted set of p-values has a significantly flatter distribution than the 'average curve', we assume that the studies were too underpowered to detect a potentially very small effect (15,16,17). If the curve is neither right-skewed nor flatter than the average powered curve, then the curve is considered too *"noisy"* to derive conclusions from it (14, 15).

In this context, we performed two p-curve analyses. The first one, named as "*p-curve*" in our report, includes all the significant p-values we found for comparisons between one intervention and one control intervention (two arms) (15, 19). As including multiple, significant p-values extracted from the same measurement in one p-curve can compromise the validity of the analysis, we allotted significant p-values that could not be included in the main p-curve to the second p-curve, the "*robust p-curve*" (15, 16). Finally, trialists sometimes reported p-values without the corresponding test-statistics, and also a few test-statistics could not be computed with the p-curve application; however, we included the tests in one of our plots to show the "whole picture", as specified in the results section below.

3. Results

3.1 Extraction results

The original meta-analysis included 41 studies, of which eleven did not report the original test statistics nor enough data to recalculate them. Six (14.6%) of these excluded articles neither reported a p-value nor a test-statistic; three (7.3%) did not make a statistical estimation of the difference between NDST and another psychotherapy at post-treatment; and, two articles (4.9%) involved Fisher's test and relative risks, which cannot be analyzed in the p-curve app. We excluded one study (2.4%) applying a non-inferiority design in which significance is interpreted differently. The *"Results"* table of one article misreported all p-values and test-statistics, so we chose not to extract further information from that study.

Thus, we included 29 (70.7%) studies in our analyses, from which we extracted a total of 43 test statistics. Seventeen (41.5%) studies reported one relevant test statistic, and twelve (29.3%) provided at least two different statistical values due to multiple distinct measures of depression.

Five statistically significant p-values turned out to be non-significant when we recalculated them. One article described two reversing interaction effects based on measurements with different instruments. We extracted four single effects from those two reversing interactions, and we included two single effects in our *"p-curve,"* while the other two were incorporated in the *"robust p-curve."*

For the p-curve analyses, we excluded another 5 (12.2%) trials because they compared three or more treatments. One (2.44%) other trial was also excluded because the authors used a *"chi-square test"* to compare three psychotherapies, essentially conducting an omnibus test. Consequently, both our *"p-curve"* and *"robust p-curve"* consisted of 23 (56.09%) trials with 24 test statistics each. Of these 24 test statistics, 13 were different for the robust analysis, as trials frequently reported multiple results for post-treatment depression. Among the analyzed 23 trials, twenty (48.78%) hypothesized that alternative psychotherapy was superior to NDST, while the remaining three (7.32%) did not.

3.2 P-curves

In **Table 1**, we present the results of our analysis of p-curves. The significant p-values included in *"p-curve"* and in *"robust p-curve"* were 13 and 14, respectively. Neither p-curve was significantly right-skewed, so we could not conclude that the findings contained evidence of effect. At the same time, neither p-curve was significantly flatter than a curve powered at 33%, so we could not conclude they lacked it either. Sensitivity analyses (see *"Appendix 1"*) showed that even after excluding extremely high or low p-values, the interpretation of the final p-curves would not change. Hence, our p-curve analyses do not depend on a few extreme test statistics.

The p-curves showed low estimations of average power: 14% for the main p-curve and 22% for the robust p-curve. The upper bounds of the 90% confidence intervals—the default confidence intervals for p-curve analysis (15)—were above 33%, so we could not reject the null hypothesis that indicated that both p-curves were flatter than one at 33% power, meaning that there was too much noise in the data to do so. In addition, the p-curves were characterized by uniform distributions (see **Figure 1**), suggesting that some results with potentially detected differences were mixed with multiple results that may could have been selectively reported.

In **Figure 2**, we show how the extracted and analyzed effect sizes, sample sizes, and p-values converge. Nine trials had a sample size between 10 and 25 participants; twelve trials had a sample size between 25 and 50; twelve trials had a sample size between 51 and 150; and eight trials had a sample size between 151 and 410 participants. While in most trials with fewer than 50 participants the results were significant ($p < .05$), less than half of the tests of differences between interventions were also significant in trials with sample sizes from 51 up to 410 participants. In this regard, the plot reflects our p-curves' estimations of low average power in the set of statistically significant results. However, **Figure 2** also shows that because trials with large sample sizes were included in the analysis, the p-curves still suggested the presence of trials with considerable power to detect very small effects.

4. Discussion

After analyzing a set of significant results from twelve clinical trials comparing NDST with other psychotherapies, our results were too noisy to allow us to make strong inferences about the presence or absence of publication bias. This may mean that there was no relation between the size of the studies, their power, and the likelihood to report a statistically significant effect.

In their report, Cuijpers et al. (11) mention that *"[t]he analyses in which we adjusted for potential publication bias… suggested that there may be publication bias"* (p.455). We have further analyzed their included trials using a different statistical method based on a distinct fundamental assumption to try to find clarity for their results. First, with p-curve analysis one can only analyze significant results ($p < .05$) because publication bias in non-significant results ($p > .05$) raises assumptions that are, at least, challenging to adopt. We agree with Simonsohn, Simmons, and Nelson when they claim that: *"while we know that the published literature generally does not discriminate among p < .05 results … we don't know how much easier it is to publish some non-significant results rather than others"* (20). When it comes to publication, they explain that the difference between a study with a p = .03 and p = .04 might be negligible (20). However, it would be difficult to assume that, for example, a p = .06 in the wrong direction is equally likely to be published as, for example, a p = .06 in the right direction (20). This assumption differentiates our approach to the publication bias problem from other statistical methods that include all possible p-values in the analysis. This assumption, however, is not flawless, and as a result there is a loss of information because of the exclusion of non-significant results.

Our approach has some limitations. Thirteen p-values included in our "*p-curve*" were different from those included in our "*robust p-curve*", because trialists reported multiple results for post-treatment depression, such as remission and response rates. We restricted our analyses to statistically significant test statistics. This may be seen as a limitation, but one or both of these significant results may have been crucial for publication.

Analyzing publication bias and estimating a pooled effect size may reasonably be seen as two different objectives when multiple results are reported. Provided that researchers in the future are interested in conducting a meta-analysis together with publication bias analysis, they should consider this problem of data collection and potentially mutually discordant objectives. After all, one result may be of interest for a meta-analysis but not for publication bias analysis, and vice versa.

Recently, p-curve analysis has been criticized as suffering from poor statistical properties, such as inadmissibility and non-monotonicity (21). These criticisms have also been debated in a blog post (22). Even so, if we discard both the p-curve and the "trim-and-fill" approach from our interpretation, relying in the results of the "limit meta-analysis" and the "step function selection model" in the original meta-analysis suggest a considerable reduction for the original effect size estimation {Hedge's g = -.21; 95% CI[-.31, -.11]}. The effect size adjustment of the former {Hedge's g = -.02; 95% CI[-.18, .013]} is larger than the latter {Hedge's g = -.09; 95% CI[-.18, .00]}, which remained statistically significant. These results would be consistent with our conjecture that large and small studies have been mixed in this meta-analysis, possibly obscuring and reducing the influence that potential publication bias could have on the final pooled effect size. However, we cannot determine the statistical uncertainty and/or magnitude of this claim, so it remains, at best, as only an educated guess.

It is reasonable to assume that clinical trials testing NDST have progressively increased their sample sizes, since almost 30 years ago it was suggested that claims of inferiority to other therapies were due to trialists not delivering NDST to patients as a treatment intended to be therapeutic because NDST was used for control groups (23, 24). This is a concrete problem in psychology, however, we can imagine that in other fields trialists face similar problems, so, as time pass by, sample sizes might be getting increasingly larger. Then, other publication bias analyses in other fields could be similar to the one we did here; a mix of trials with high and low power for which publication bias test results can only be inconclusive.

The most important limitation of our study is that only one author (G.A.M.F.) initially extracted the test statistics of the studies. To counterpose this limitation, 20% of the included studies were randomly reviewed by the other two authors (A.F.M. and M.L.), and the remaining author (L.C.) checked the whole dataset afterwards. We also included discrete tests, which can serve as source of liability for p-curve analysis, even though such inclusion should negligibly affect our results (25). Also, our study is dependent on the investigation conducted by Cuijpers *et al.* (11), so it might inherently suffer from the same limitations as that study. That is, we have included trials that were originally collected by their study, so our study relies on their inclusion criteria and research process that has been documented in their article (11) and/or elsewhere (26).

5. Conclusions

Ultimately, the results of our study were considered *"too noisy"*; as a result, strong inferences about publication bias could not be drawn in this set of articles. We conclude that there were indications of significant results from small-sample trials that could have been due to publication bias; and that there were trials with large samples that might have been able to detect potential treatment differences. Possibly due to this reason, our results suggested that the complete set of trials had a low average statistical power.

We have attempted to show how and what it would entail to apply a distinct publication bias method such as *"p-curve"* in clinical trials. Hence, this report has been written from the perspective of empirical researchers examining a challenging *"case study"* to show the assumptions and analytic choices underlying publication bias analysis. In so doing, we have highlighted how complex publication bias analysis is, hoping that this report could serve as a valuable tool for other researchers to reflect when dealing with publication bias analysis from any kind of method.

Declarations

Ethics and Consent to Publish: Not applicable.

Availability of data and materials: All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Competing Interests: The authors declare that they have no competing interests

Funding: No external funding was received for this study.

Author contributions: GAMF conceptualized the study, developed the methodology, and analyzed the data. LC examined the validity and reliability of the data, mediated in regards to analytical choices, and reviewed the final manuscript. AFAM examined the validity and reliability of part of the data, and he reviewed the final manuscript as well as the appendices. ML supervised the study, together with examining the validity and reliability of part of the data, and she reviewed the final manuscript as well

as the appendices. GAMF wrote the first draft of the manuscript, while ML and AFM edited and critically revised several parts of it.

Acknowledgements: We are indebted to Dr. Pim Cuijpers, who shared his database with us.

References

1. Dickersin K. The Existence of Publication Bias and Risk Factors for Its Occurrence. JAMA. 1990;263(10):1385-9.
2. Muradchanian J, Hoekstra R, Kiers H, Van Ravenzwaaij D. The role of results in deciding to publish: A direct comparison across authors, reviewers, and editors based on an online survey. PLOS ONE. 2023;18(10):e0292279.
3. Sterling TD, Rosenbaum WL, Weinkam JJ. Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. The American Statistician. 1995;49(1):108-12.
4. Fanelli D. Negative results are disappearing from most disciplines and countries. Scientometrics. 2012;90(3):891-904.
5. Friese M, Frankenbach J. p-Hacking and publication bias interact to distort meta-analytic effect size estimates. Psychological methods. 2020;25(4):456-71.
6. Lash TL. The Harm Done to Reproducibility by the Culture of Null Hypothesis Significance Testing. Am J Epidemiol. 2017;186(6):627-35.
7. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56(2):455-63.
8. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. Bmj. 1997;315(7109):629-34.
9. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011; 343(7818):1-8.
10. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ. 2006;333(7568):597-600.
11. Cuijpers P, Miguel C, Ciharova M, Harrer M, Karyotaki E. Non-directive supportive therapy for depression: A meta-analytic review. Journal of Affective Disorders. 2024;349:452-61.

12. Rücker G, Schwarzer G, Carpenter JR, Binder H, Schumacher M. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. Biostatistics. 2011;12(1):122-42.
13. McShane BB, Böckenholt U, Hansen KT. Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. Perspectives on Psychological Science. 2016;11(5):730-49.
14. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. Doing Meta-Analysis With R: A Hands-On Guide. 1st ed. Boca Raton, FL and London: Chapman & Hall/CRC Press; 2021 2021.
15. Simonsohn U, Nelson LD, Simmons JP. P-curve: A key to the file-drawer. Journal of Experimental Psychology: General. 2014;143(2):534-47.
16. Simonsohn U, Nelson LD, Simmons JP. P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. Perspectives on Psychological Science. 2014;9(6):666-81.
17. Simonsohn U, Simmons JP, Nelson LD. Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). Journal of Experimental Psychology: General. 2015;144(6):1146-52.
18. Simonsohn U, Nelson LD, Simmons JP. P-curve won't do your laundry, but it will distinguish replicable from non-replicable findings in observational research: Comment on Bruns & Ioannidis (2016). PloS one. 2019;14(3):e0213454.
19. Nelson L, Simonsohn U. [60] Forthcoming in JPSP: A Non-Diagnostic Audit of Psychological Research. Data Colada. (2020, February 12). Available from: https://datacolada.org/60
20. Simonsohn U, Simmons J, Nelson L, [61] Why p-curve excludes ps>.05. Data Colada. (2017, June 15). Available from: https://datacolada.org/61
21. Morey RD, Davis-Stober CP. On the poor statistical properties of the P-curve meta-analytic procedure. Journal of the American Statistical Association.1-19.
22. Simonsohn U. [129] P-curve works in practice, but would it work if you dropped a piano on it? (2025, September 23). Available from: https://datacolada.org/129

23. Wampold BE, Mondin GW, Moody M, Stich F, Benson K, Ahn H-n. A meta-analysis of outcome studies comparing bona fide psychotherapies: Empiricially, "all must have prizes.". Psychological Bulletin. 1997;122(3):203-15.

24. Wampold BE, Imel ZE, Miller SD. Barriers to the Dissemination of Empirically Supported Treatments: Matching Messages to the Evidence. Behavior Therapist 2009;32(7):144-55.

25. Simonsohn U, Nelson LD, Simmons JP. Supplementary Materials P-curve: A Key to The File Drawer. P-curve.com. (2015, September 22). [Accessed on: 2025, 05, 22] Available from: https://p-curve.com/Supplement/Supplement_pcurve1.pdf

26. Cuijpers, P., 2023, June 14. The effects of non-directive counseling for depression: A meta-analytic review. OSF Registries. https://osf.io/rfv67

Tables and Figures.

Table 1. Results of the calculated p-curves.

Figure 1. P-curve plot.

Figure 2. Extracted p-values, effect, and sample sizes of the trials.

Appendix 1.

Figure A1.1. Plots on the influence of removing outliers from the p-curve.

Figure A1.2. Plots on the influence of removing outliers from the robust p-curve.

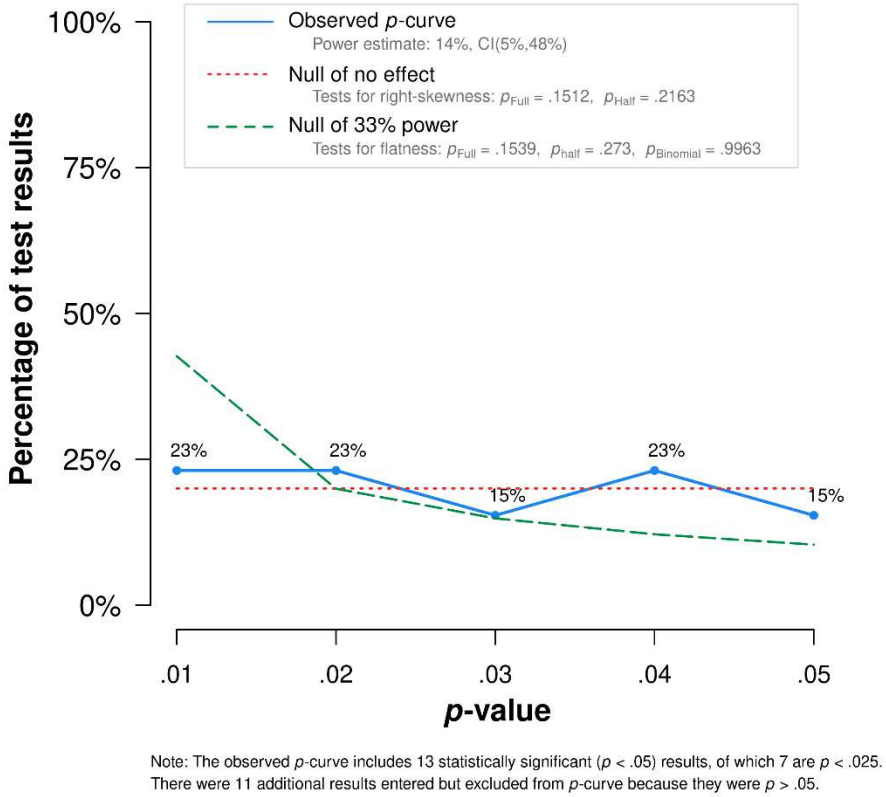
Table 1. Results of the multiple calculated *p*-curves.

	Binomial Test	Continuous Test (Aggregate with Stouffer Method)					
		Full Curve		Half Curve		Evidence of effect	
	p < .025	z-Full	p's < .05	z-Half	p's < .025	Present	Absent
				P-curve			
Right-Skewness Test	p = .5	z = -1.03	p = .15	z = -.78	p = .22	No	No
Flatness Test	p = .99	z = -1.02	p = .15	z = -.60	p = .27		
Statistical Power				Estimate: 14%			
				90% Confidence interval: (5%, 48%)			
				Robust p-curve			
Right-Skewness Test	p = .21	z = -1.61	p = .05	z = -1.07	p = .14	No	No
Flatness Test	p = .99	z = -.55	p = .29	z = -.46	p = .32		
Statistical Power				Estimate: 22%			
				90% Confidence interval: (5% , 57%)			

Note: The Stouffer Method is used to compile and analyze p-values. P-curve can be divided by its full distribution, i.e., from .05 toward the non-central distribution, and its half distribution, i.e., from .025 toward the non-central distribution. Evidence of an effect (absence of selective reporting) is assumed when both full and half distributions are significantly right-skewed at p < .1. or when only half of the distribution is significantly right-skewed at p < 0.5. None of the p-curves contain evidence of an effect. When the full p-curve is flatter (p < .05) than one powered at 33%, or when both the 33% binomial test and half of the p-curve are statistically significant (p < 0.1), then selective reporting cannot be ruled out, which is the case for both p-curves.

[*p < 0.5]

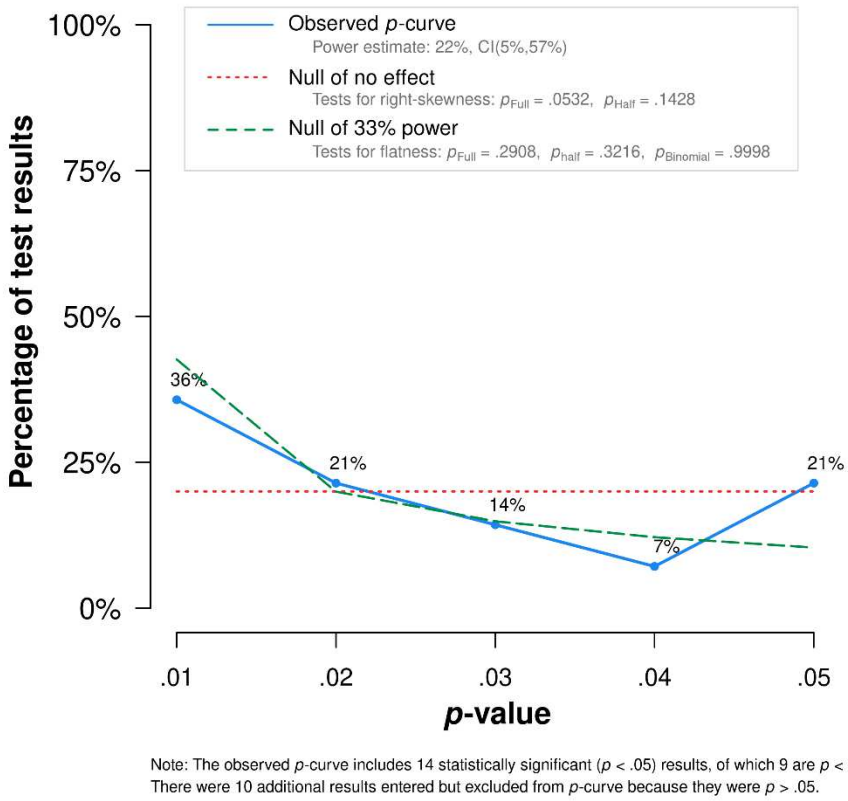
Figure 1. P-curve plot



Note: The observed *p*-curve includes 13 statistically significant ($p < .05$) results, of which 7 are $p < .025$.
There were 11 additional results entered but excluded from *p*-curve because they were $p > .05$.

P-curve of 23 trials containing 24 p-values; 13 are statistically significant. The blue line shows the distribution of critical p-values: 15% of the p-values fall between .04 and .05; 23% between .03 and .04; 15% between .02 and .03; 23% between .01 and .02; and 23% between .00 and .01. The text box displays the average power and its confidence interval for the included studies. The red line represents the distribution when there should be no effect. The green line shows what the distribution would look like if a group of studies were powered at the 33% arbitrary threshold to determine whether an effect was non-negligible. By observing the curve, selective reporting can neither be ruled in or ruled out. The uniform shape suggests a group of potentially selectively reported findings mixed with some that estimated differences. This figure was generated by the p-curve app 4.10 (https://www.p-curve.com/app4/).

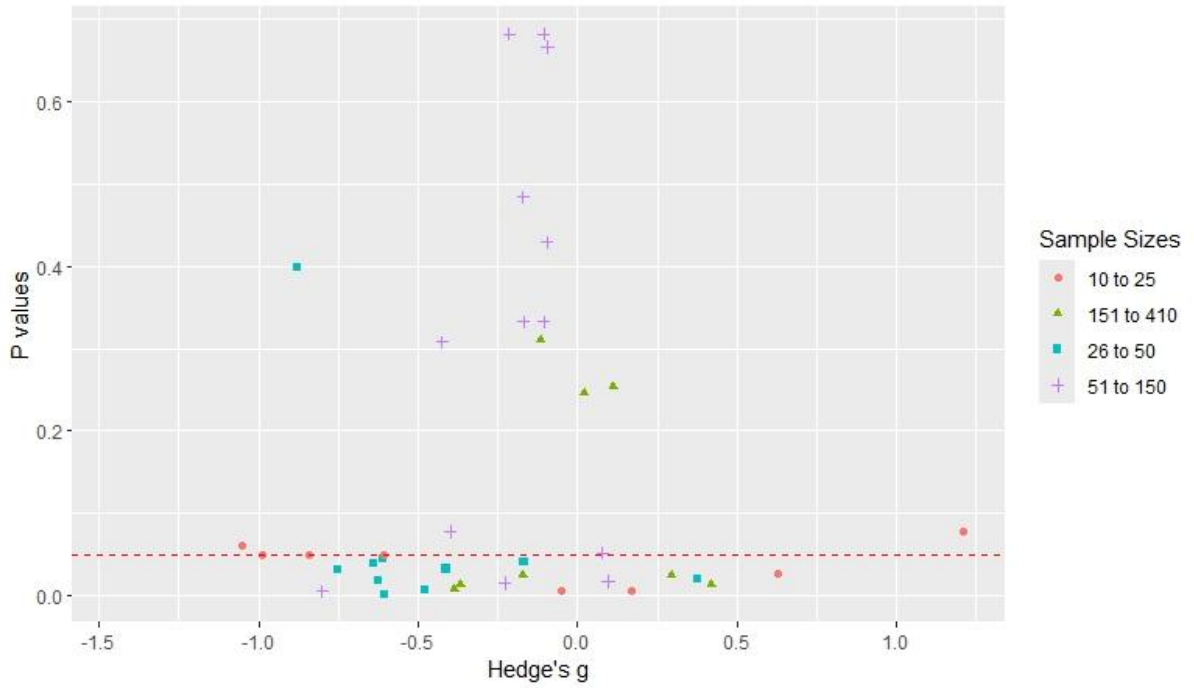
Figure 1b. Robust p-curve plot



Note: The observed *p*-curve includes 14 statistically significant ($p < .05$) results, of which 9 are $p < .025$.
There were 10 additional results entered but excluded from *p*-curve because they were $p > .05$.

Robust p-curve of 23 trials containing 24 test statistics; 14 are statistically significant. Full-curve right-skewness: z = -1.61, p = .05; half-curve: z = -1.07, p = .14. The curve is not flatter than one powered at 33% (p = .30). The robust p-curve is not statistically significant either, and it does not formally meet the criterion for absence of selective reporting.

Figure 2. Extracted p-values, effect, and sample sizes of the trials



406

407 Plot of the extracted p-values, effect, and sample sizes from the different trials. The red dotted line is

408 the significance threshold ($p < .05$). The *"y-axis"* represents the p-values, in which the higher ones are

409 at the top, while the x-axis depicts the values of the effect sizes—the left favors other psychotherapies

410 and the right NDST. Fifteen effect estimations (83.33%) with sample sizes ranging from 10 to 50 were

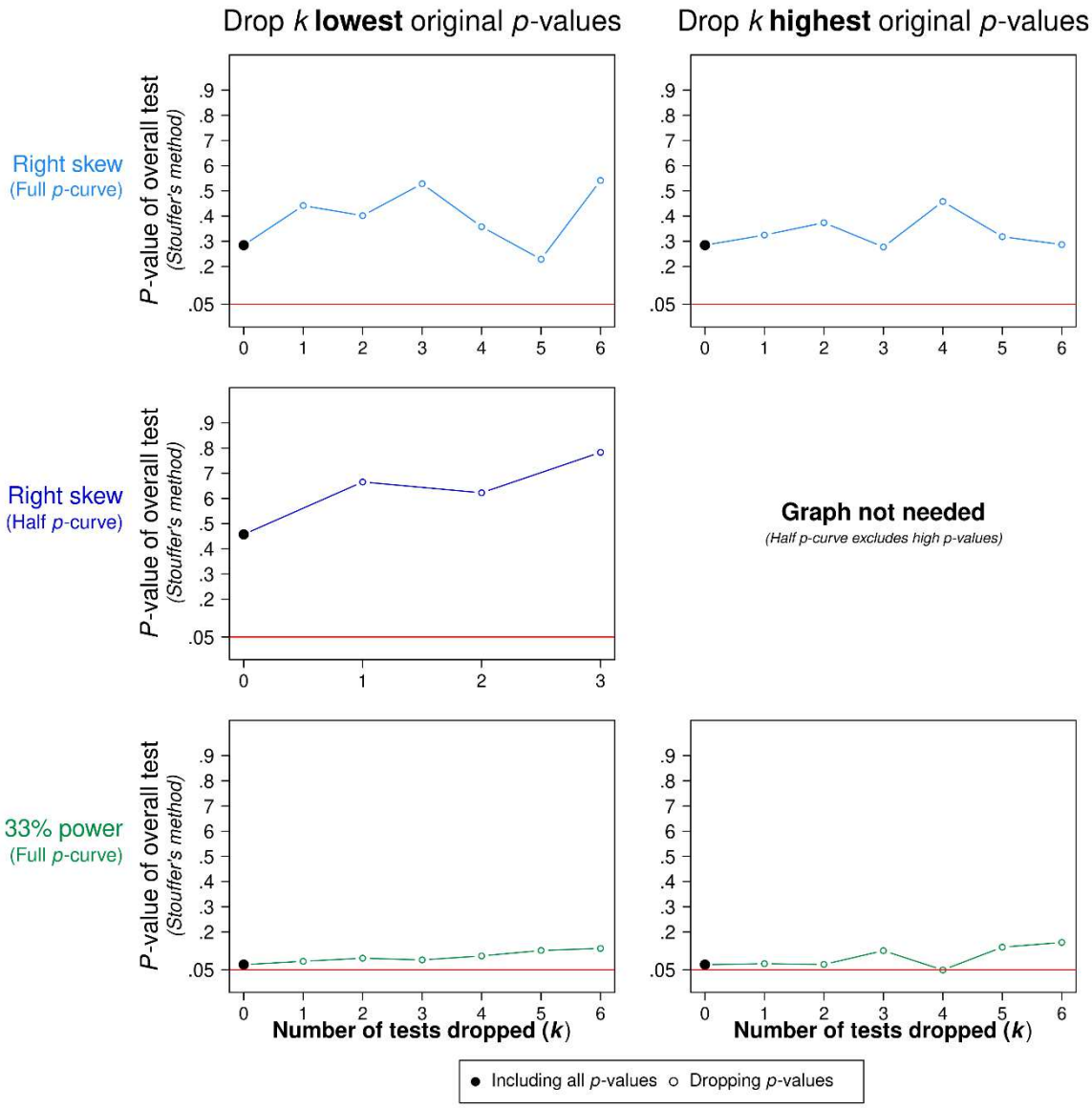
411 significant, while twelve effect estimations (57.14%) with more than 51 participants exceeded the

412 significance threshold.

413

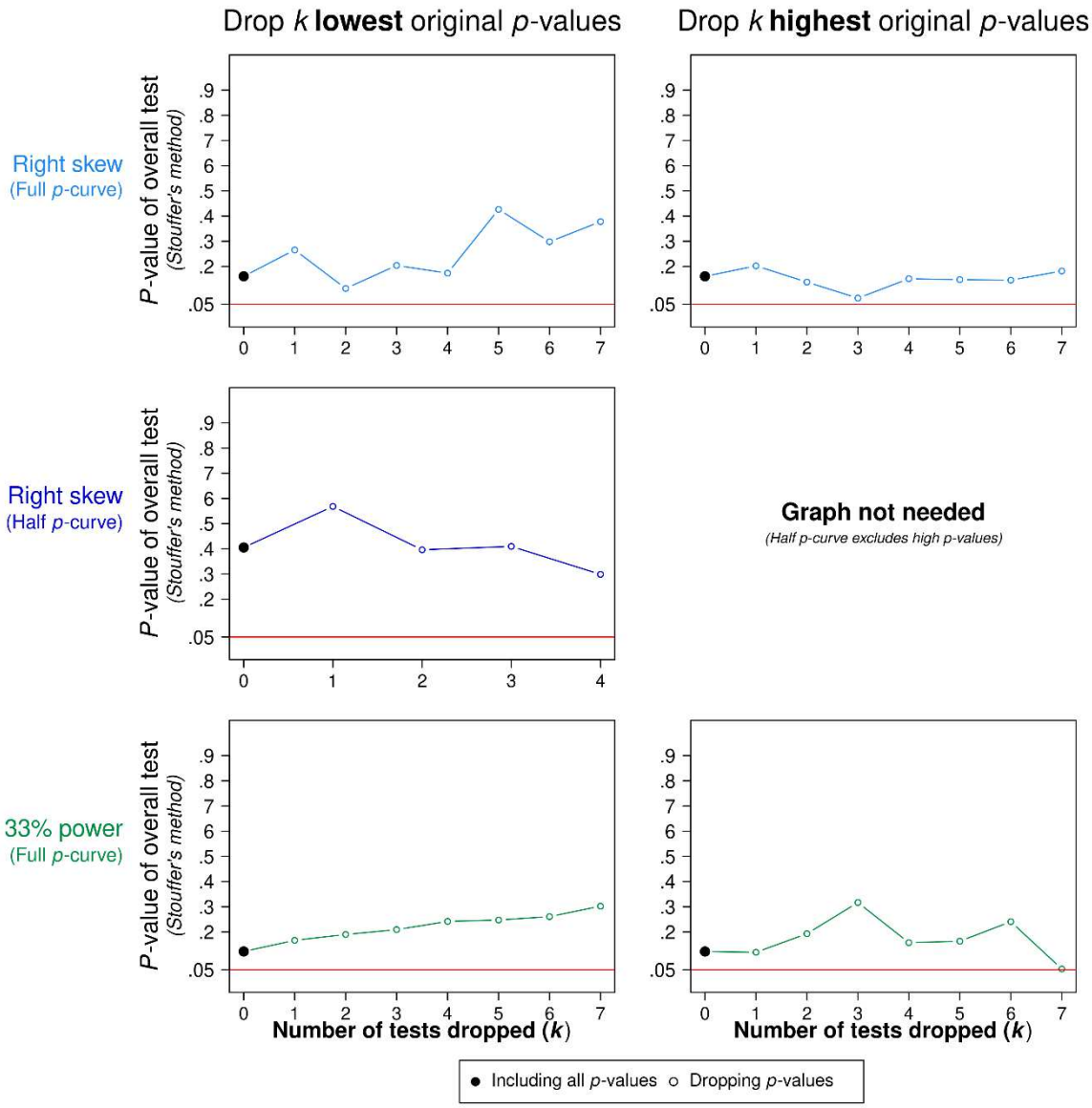
Appendix 1.

Figure A1.1. Plots on the influence of removing outliers from the p-curve.



The first three plots show that neither dropping the highest nor the lowest p-values would make the p-curve to be right-skewed and, hence, to contain evidential value. Based on the two plots at the lower part of the figure, dropping multiple options of the lowest and highest p-values would not change the distribution, which would still not be flatter than one powered at 33%.

Figure A1.2. Plots on the influence of removing outliers from the robust *p*-curve.



The first three plots show that neither dropping the highest nor the lowest p-values would make the *p*-curve to be right-skewed and, hence, contain evidential value. Based on the two plots at the lower part of the figure, dropping multiple levels of the lowest and highest p-values would not change the distribution, which would still not be flatter than one powered at 33%.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.xlsx