Replication Crisis and Its Solutions

Esa Palosaari <esa.palosaari(at)uta.fi>

University of Tampere

20.11.2017

Contents

1. Practical matters
2. Background
3. Reproducibility projects
4. Interpreting replication failure
5. Improving statistical inferences
   - *P*-curve analysis
   - Preregistration

Practical matters

Timetable

10:15 – 12:00	Interpreting replication failures
12:00 – 13:00	Lunch
13:00 – 14:00	Introduction to *p*-curve analysis and preregistration
14:15 – 16:00	*P*-curve analysis and preregistration excercises

Course website

users.aalto.fi/~palosae2/

Evaluation

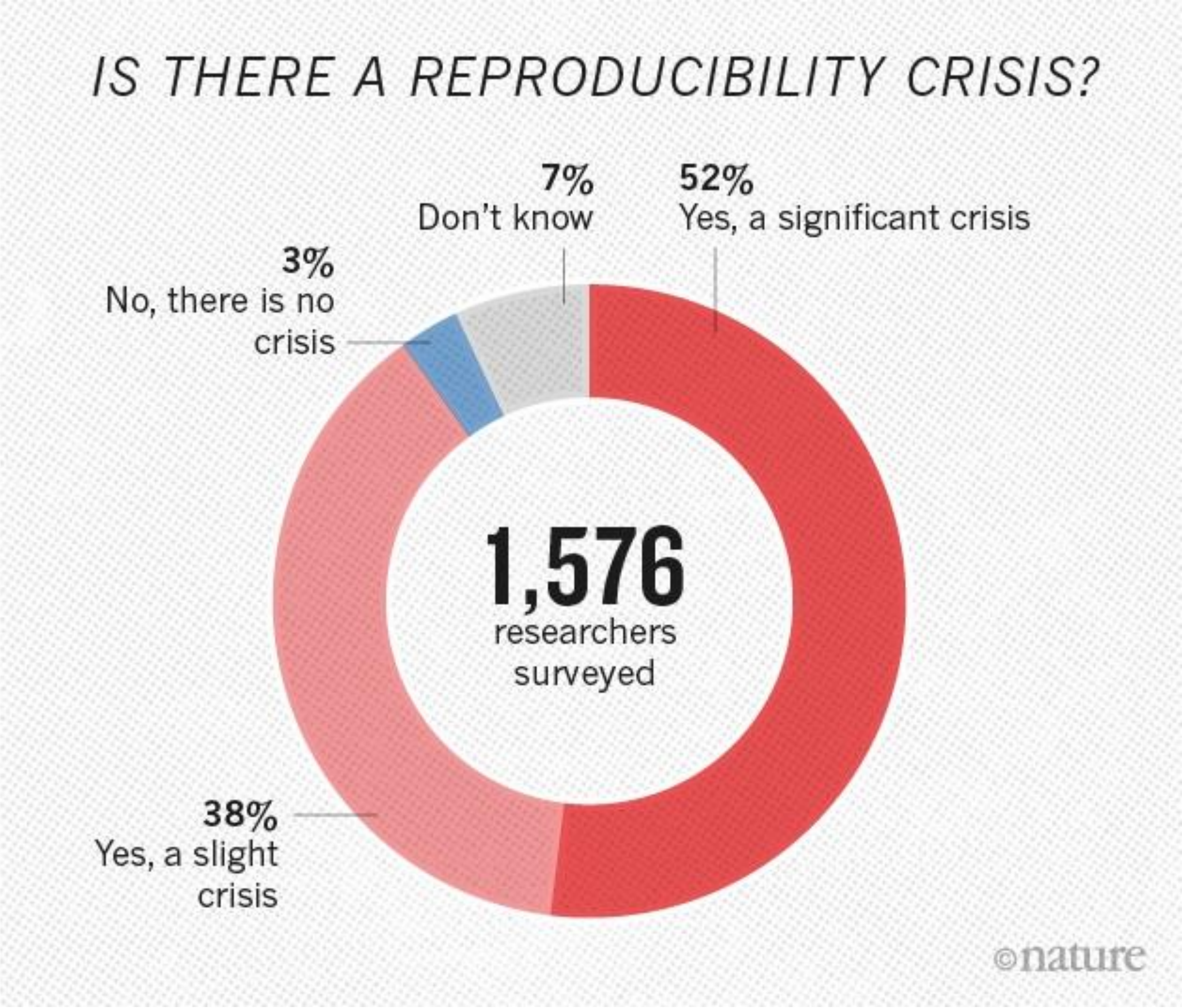
Pass/fail based on attendance and two excercises.

Send a *p*-curve analysis and a link to a preregistration to esa.palosaari(at)uta.fi by 27.11.

Polls

Please respond to the anonymous in-class polls at PollEv.com/esapalosaari182

Nature’s online survey



Background: Personal

Who am I to talk about these things?

Just another researcher

Ph.D. in Psychology from UTA in 2016

- *P*-value crisis while submitting thesis

Reading group hardsci.wordpress.com/2016/08/11/everything-is-fucked-the-syllabus/

One ongoing preregistered experimental study, including a *p*-curve analysis

Still trying to figure these things out myself…

Background: Why do reproducibility projects?

”Scientific claims should not gain credence because of the status or authority of their originator but by the **replicability of their supporting evidence**”

”**Even research of exemplary quality may have irreproducible empirical findings** because of random or systematic error.”

”Practices and incentives [--] may inflate false-positive [--] or irreproducible results. Potentially **problematic practices include *selective reporting, selective analysis*, and insufficient specification of the conditions** necessary or sufficient to obtain the results.”

(Open Science Collaboration, 2015)

Background: Unreliable literature?

Concerns that the published literature is biased and irreproducible partly because of

1. **Publication bias (*selective reporting*)**
   Favoring the publication of statistically significant findings

2. **Unreported flexibility in data analysis (*selective analysis*)**
   Allows almost any result to become significant
   The garden of forking paths
   Hypothesising After Results are Known (HARKing)

Interlude: *P*-values

Let’s run 100 000 experiments about the same question

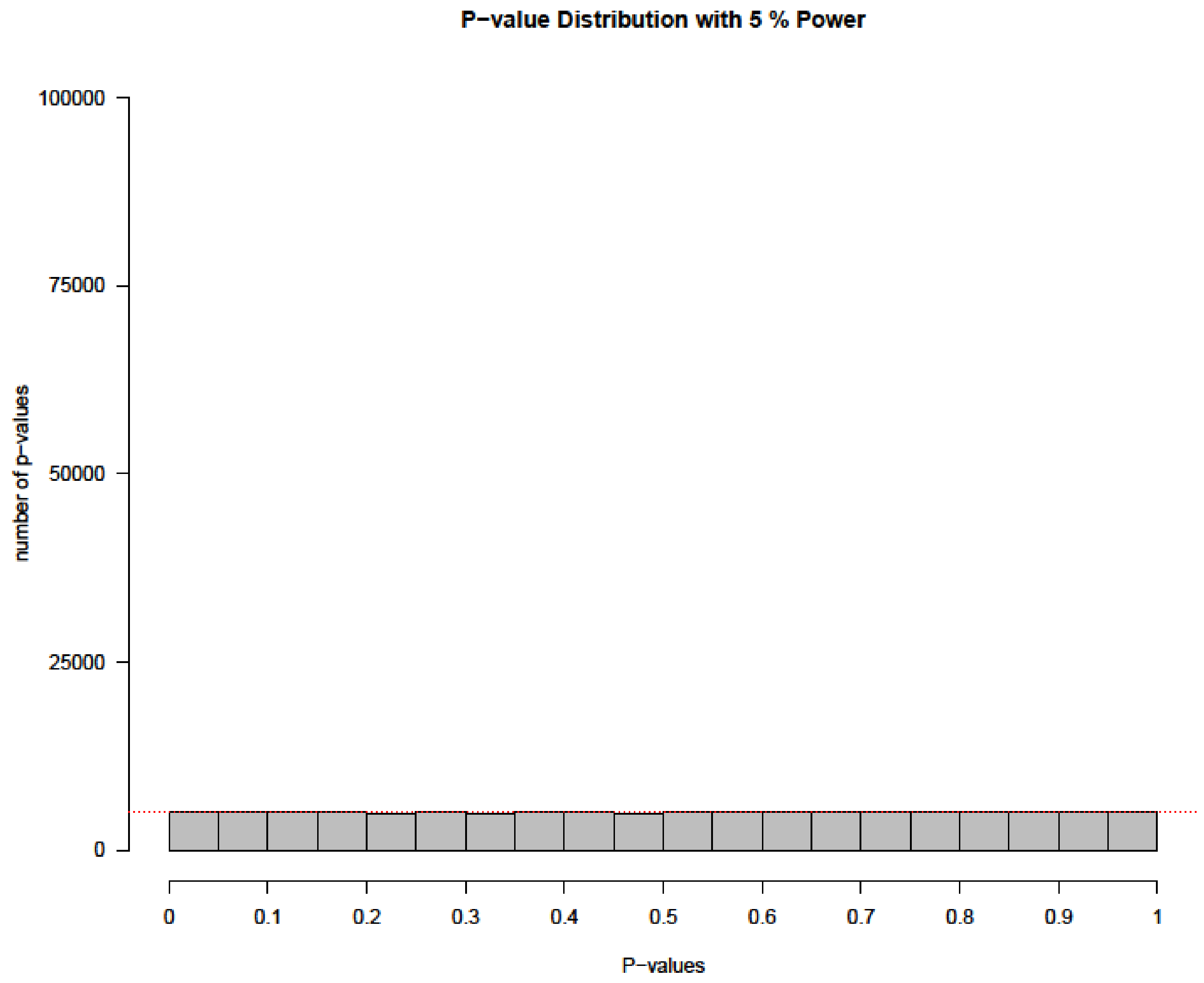
Two groups, sample size 20 for each

First, let’s set the true group means to be equal

What *p*-values can you expect?

- Please, do not look ahead in the slides

Poll https://PollEv.com/esapalosaari182



R code: https://users.aalto.fi/~palosae2/pvaluesTwoSample0.R

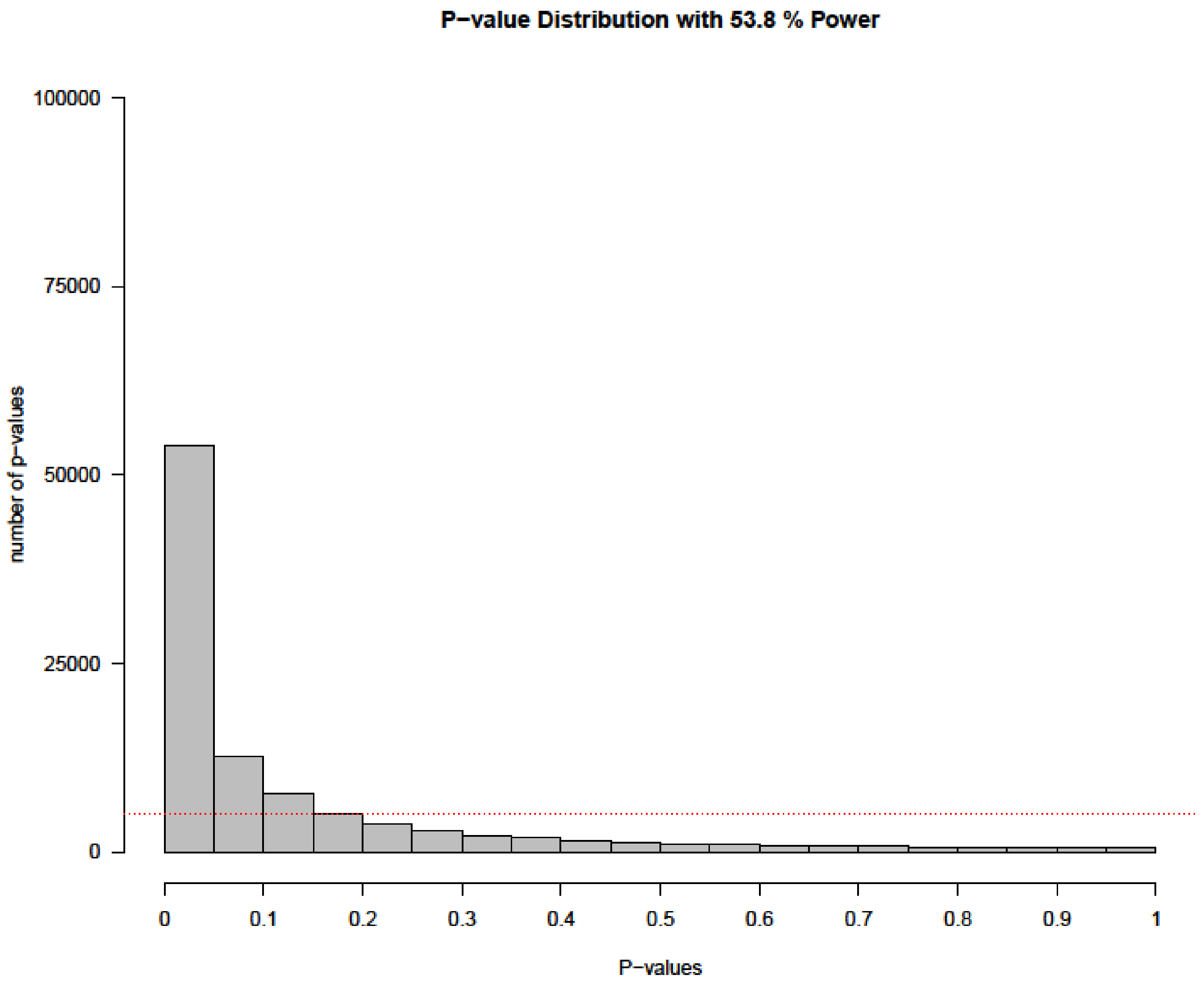
Interlude: *P*-values

Let’s set the true difference between group means to 10

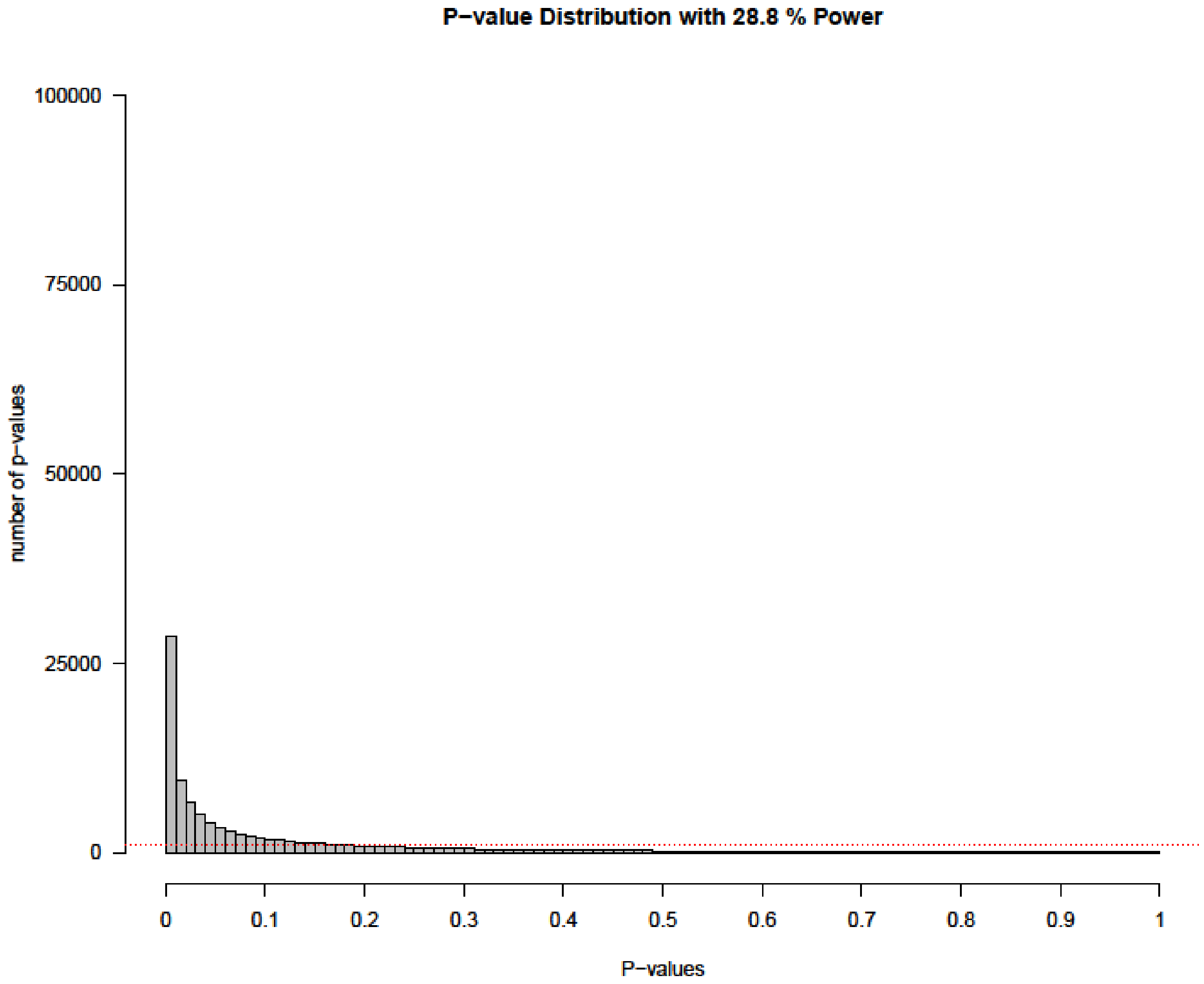
Both groups have a standard deviation of 15

Both groups have a sample size of 20

Poll https://PollEv.com/esapalosaari182

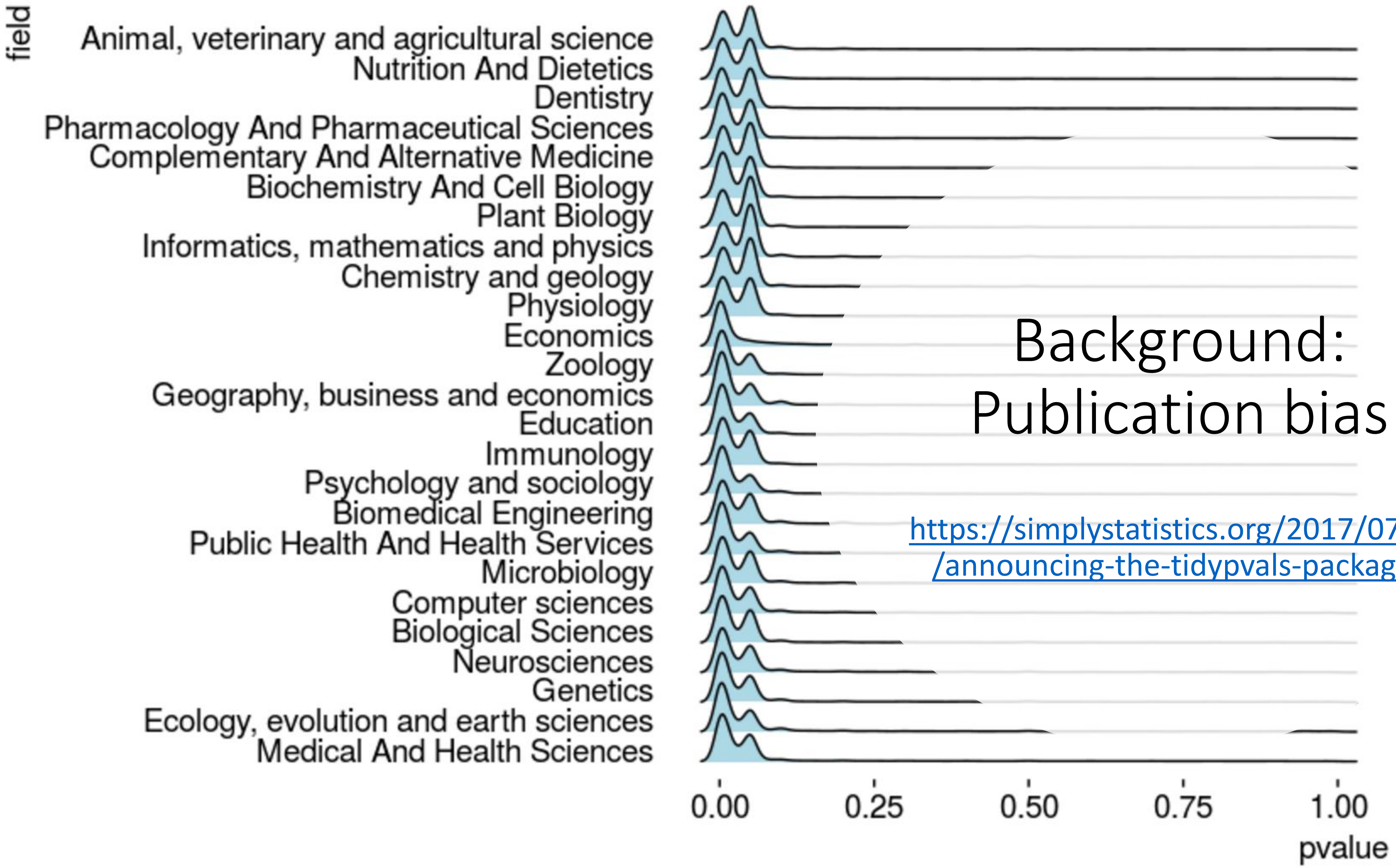


R code: https://users.aalto.fi/~palosae2/pvaluesTwoSample1.R



R code: https://users.aalto.fi/~palosae2/pvaluesTwoSample2.R

Background: Publication bias

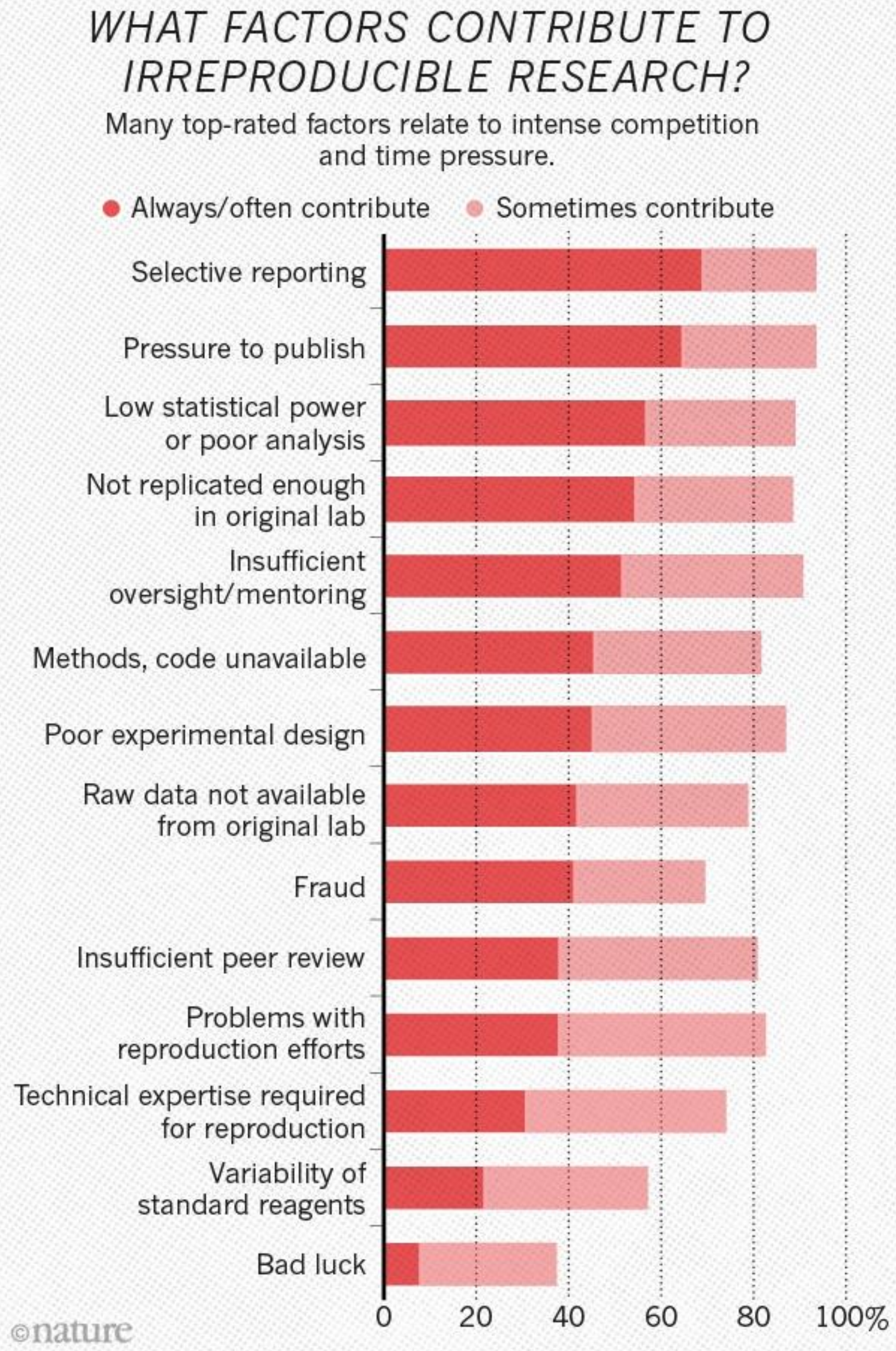


https://simplystatistics.org/2017/07/26/announcing-the-tidypvals-package/

Background: Publication bias

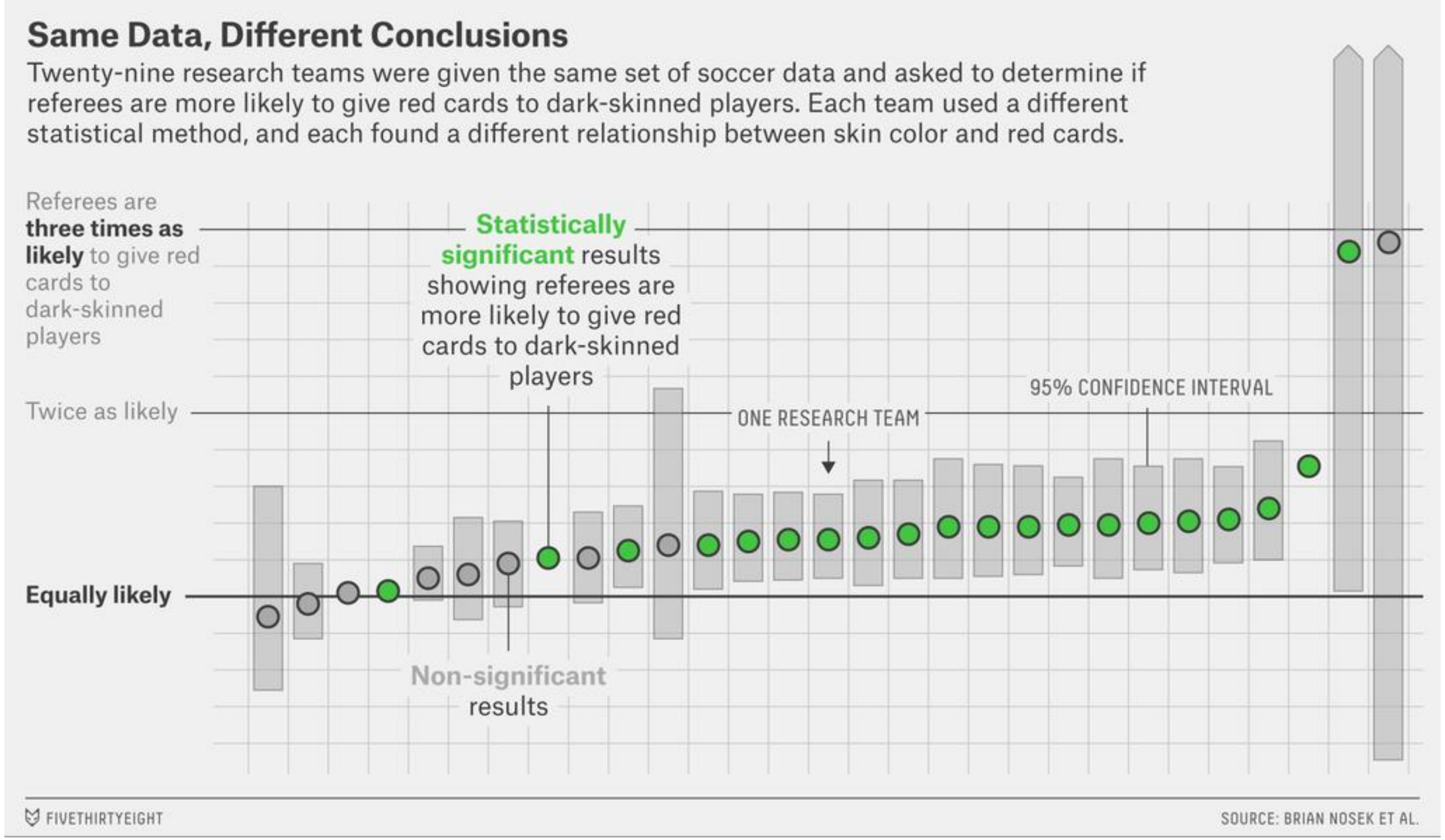
In biomedical research, pre-study plans or protocols are commonly required by ethics committees or by the state

Study	Sample	Significant results more likely to be published
Dickersin & Min, 1993	198 NIH trials funded in 1979	OR = 12.3, 95% CI [2.5, 60.0]
Chan & Altman, 2005	All PubMed articles in 2000	OR = 2.0, 95% CI [1.6, 2.7]
Decullier & Chapuis, 2005	649 French protocols	OR = 4.6, 95% CI [2.2, 9.5]
Song & al., 2009	12 studies about publication bias	OR = 2.8, 95% CI [2.1, 3.7]



Background: Publication bias

Background: Flexibility in data analysis



Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

SOURCE: BRIAN NOSEK ET AL.

Study: https://osf.io/j5v8f/

Figure: https://fivethirtyeight.com/features/science-isnt-broken/#part1

Background: Flexibility in data analysis

The garden of forking paths (Gelman & Loken, 2013)

= data analysis dependent on the data at hand,

rather than analysis rules being prespecified before seeing the data.

That is: **exploratory** data analysis or ”testing” rather than **confirmatory** analysis.

With a different realization of the data set, analysis choices would have been different.

P-values are based on what would have happened in other data sets. They are inaccurate or inapplicable in exploratory analysis dependent on a specific data-set.

Problematic even if people are not ”actively trying out different tests in a search for statistical significance” (p-hacking, fishing)

Background: Flexibility in data analysis

Example of the garden of forking paths (Petersen et al. 2013 cited in Gelman & Loken, 2013)

Petersen et al. 2013 ”claimed to find an association between men’s upper-body strength, interacted with socioeconomic status, and their attitudes about economic redistribution.” No preregistered analysis plan.

Reported a statistically significant interaction, with no statistically significant main effect. That is: they did **not** find that men with bigger arm circumference had more conservative positions.

But that correlation of arm circumference with redistribution opinions was higher among men of higher socioeconomic status.

It is likely that if there was a main effect, they would have claimed that it supported their hypothesis. The same if another interaction would have been significant in this particular sample.

There are multiple analysis paths to a significant result

-> a multiple comparison problem of all the paths that could have been taken.

Replication projects

Estimating the reproducibility of fields as a whole

Samples of published studies

Direct replications rather than conceptual replications

Preregistered study designs and analysis plans

Aspirations for high statistical power

All results published regardless of statistical significance

Replication projects: Cognitive and Social Psychology

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251). (http://osf.io/ezcuj)

100 experimental and correlational studies published in three psychology journals during 2008

Psychological Science

Journal of Personality and Social Psychology

Journal of Experimental Psychology: Learning, Memory, and Cognition

Quasi-random sampling

Replication projects: Cognitive and Social Psychology

Last experiment from each article was selected

A key result was identified from the selected experiment to be replicated

Analyses for each replication study was reproduced by another independent analyst

Different standards for evaluating replication success

- Significance
- *p*-values
- Effect sizes
- Subjective assessments
- Meta-analyses of effect sizes

Replication projects: Cognitive and Social Psychology

97 of the original 100 had a $p < .05$

Expectation of 89 significant replication results if all original effects were true and accurately estimated

Only 35 were statistically significant at $p < .05$ [95% CI = (27%, 46%)]

Replication projects: Cognitive and Social Psychology

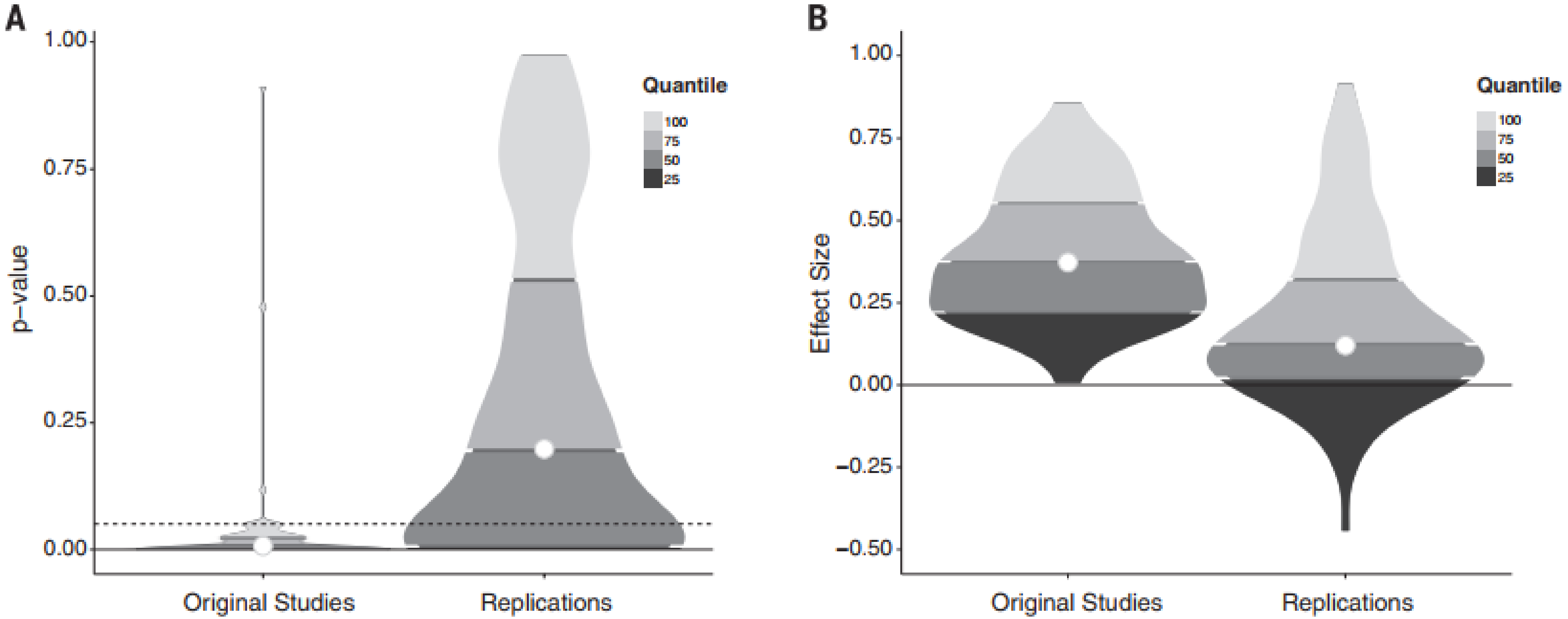


Fig. 1. Density plots of original and replication *P* values and effect sizes. (**A**) *P* values. (**B**) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

Replication projects: Cognitive and Social Psychology

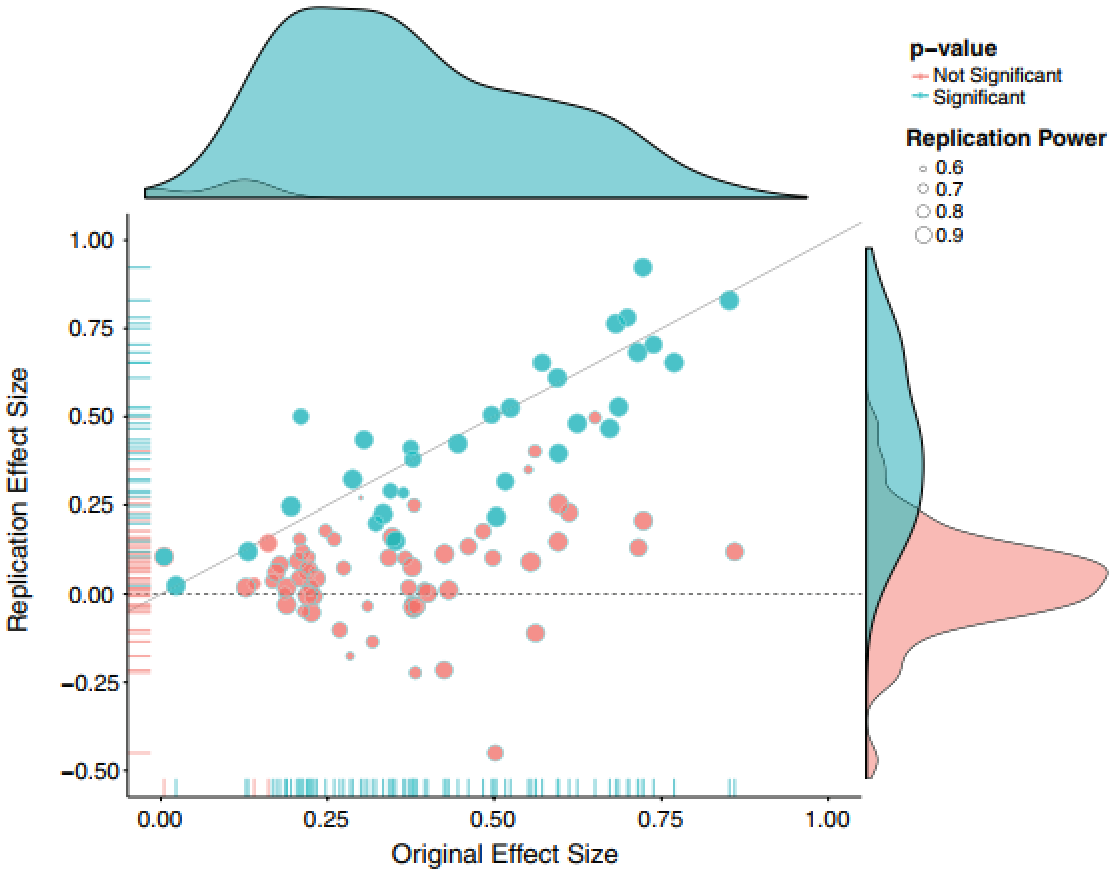


Fig. 3. Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Replication projects: Many Labs

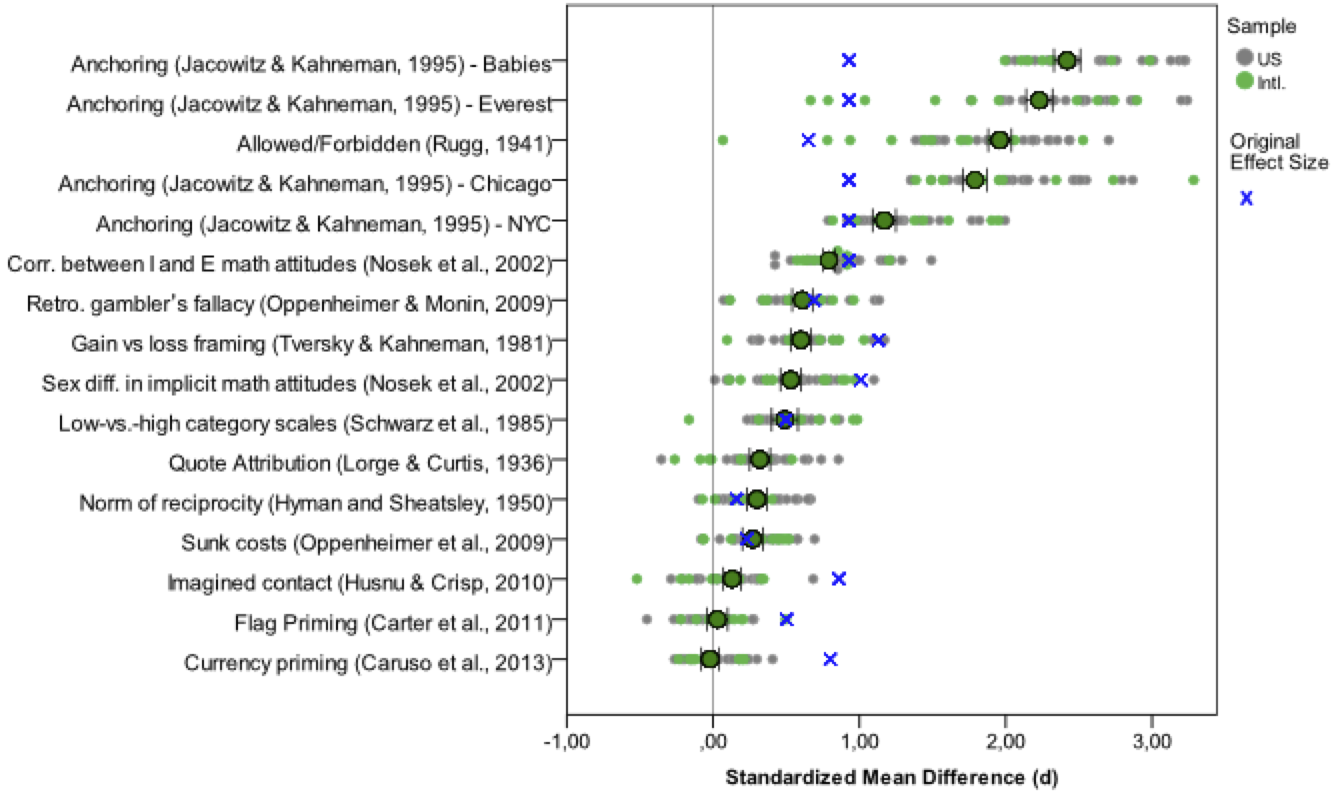
Many replications of single effects (https://osf.io/wx7ck/)

Replications of 13 classic and contemporary effects in psychology with 36 samples and 6344 participants

Original studies were **not** selected randomly (https://osf.io/3467b/)

10/13 effects replicated consistently

Replication projects: Many Labs



Replication projects: Experimental Economics

Camerer, C. F. & al. (2016). Evaluating replicability of laboratory experiments in economics. Science, aaf0918. DOI: 10.1126/science.aaf0918

Sample: all 18 between-subject laboratory experimentla papers published in the *American Economic Review* and the *Quarterly Journal of Economics* between 2011 and 2014

Most significant finding emphasized by authors chosen for replication

90% power to detect original effect size at the 5% significance

2/18 of the originals and 7/18 of the replications had $p > .05$

Replication projects: Experimental Economics

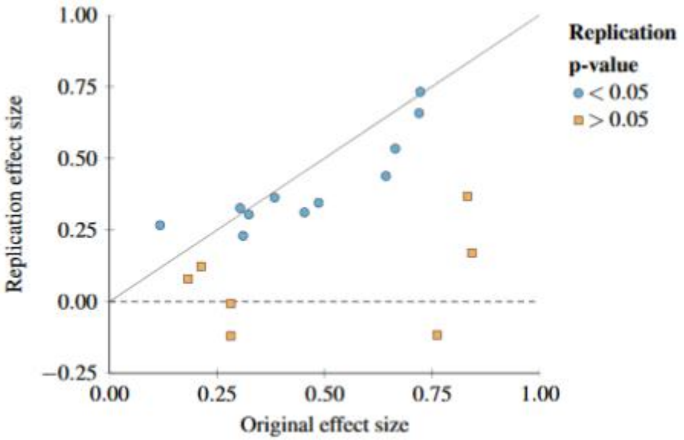


Fig. S3. Original study effect size versus replication effect size (correlation coefficients r).

The diagonal line represents replication effect size equal to the original effect size and the dotted line represents a replication effect size equal to zero. Blue dots are the replications that were significant with P<0.05 in the original direction, and red dots are the replications that were not significant. The mean standardized effect size (correlation coefficient, r) of the replications is 0.279 (SD=0.234), compared to 0.474 (SD=0.239) in the original studies. This difference is significant (Wilcoxon signed-ranks test, n=18, z=-2.98, P=0.003). The mean relative effect size of the replications is 65.9% [95% CI=(37.2%, 94.7%)]. The Spearman correlation between the original effect size and the replication effect size is 0.48 (P=0.043).

Replication projects: Cancer Biology

https://osf.io/e81xl/wiki/home/

Ongoing

Random sampling from the most cited papers in 2010 (584), 2011 (548) and 2012 (543) resulting in 50 studies to be replicated

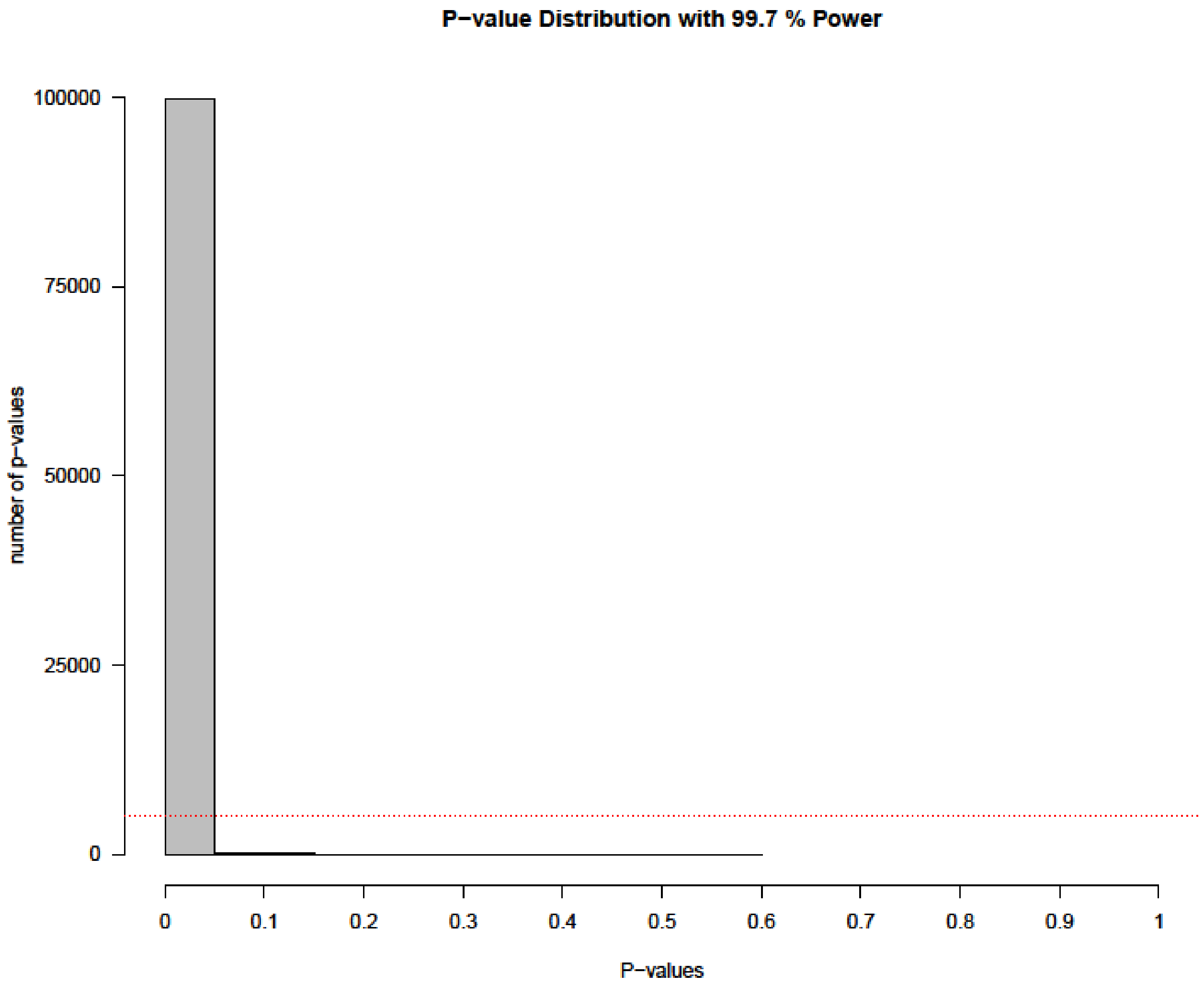
Earlier, non-transparent replication studies by two industrial laboratories

1) 6 out of 53 landmark studies replicated, 11% (https://www.nature.com/articles/483531a#t1)

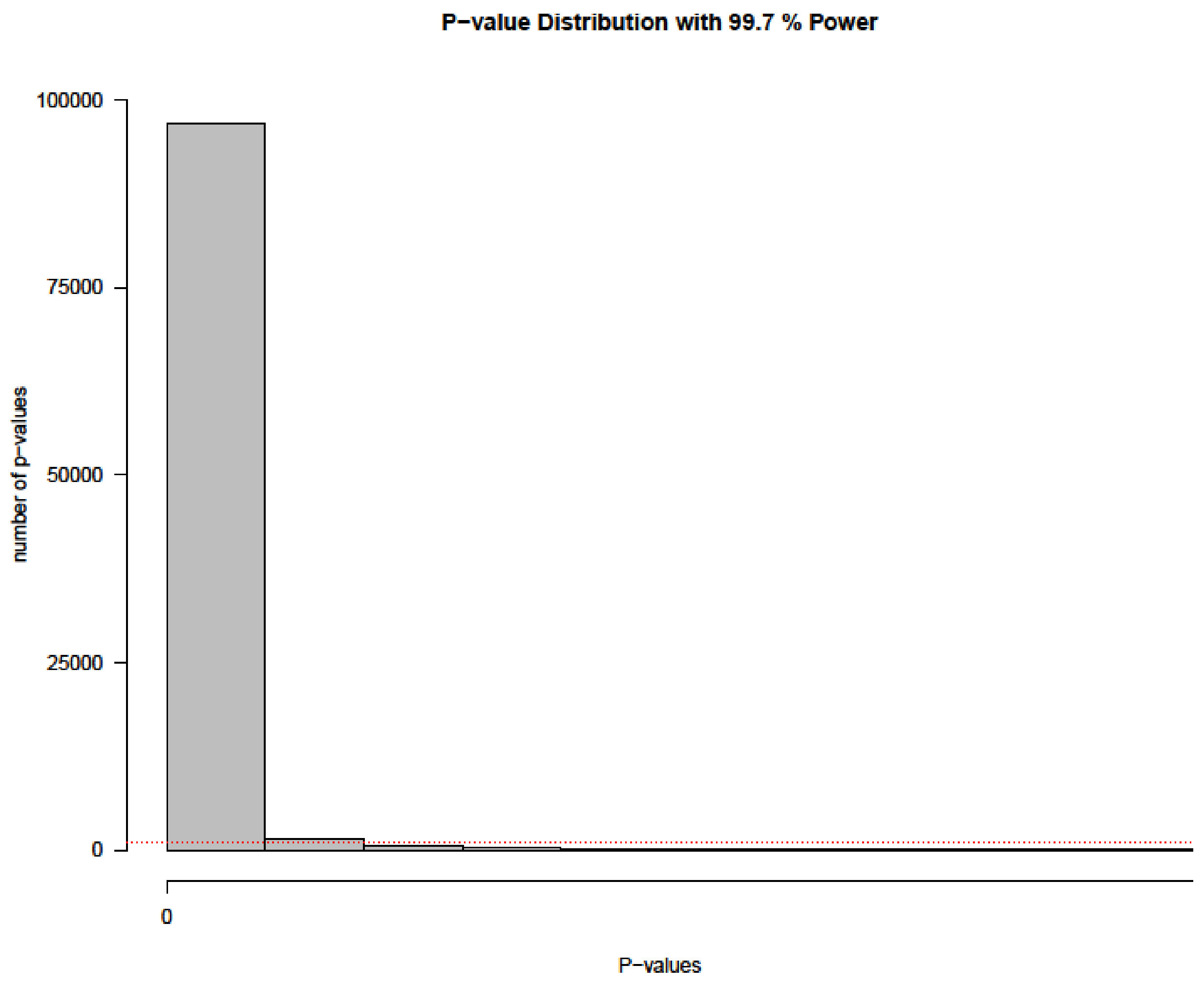
2) 20-25% replication success from 67 projects (https://www.nature.com/articles/nrd3439-c1)

Interlude: *P*-values, again

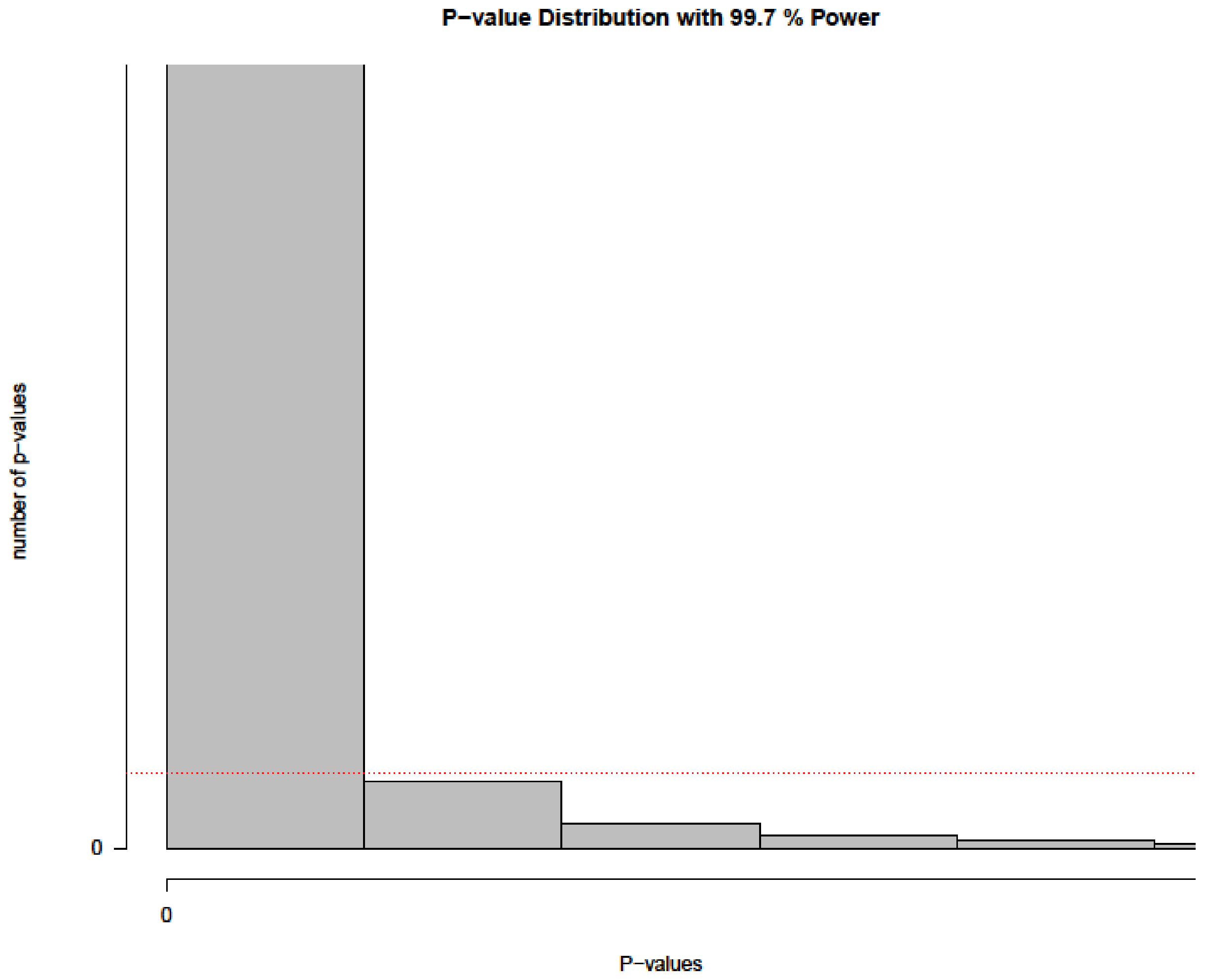
Poll https://PollEv.com/esapalosaari182



R code: https://users.aalto.fi/~palosae2/pvaluesTwoSample3.R



R code: https://users.aalto.fi/~palosae2/pvaluesTwoSample4.R



R code: https://users.aalto.fi/~palosae2/pvaluesTwoSample5.R

Interpreting non-significance: Fisher

P-values are the theoretical probability of the results under the null hypothesis (H_0)

≈ $P(D| H_0)$

Levels of significance are approximate (.049 ≈ .051) and graded

Small *p*-values are taken as evidence against the null hypothesis

Non-significant results should be mostly just ignored

”Fisher denied that the null hypothesis could ever be established but conceded that non-significant results might be used for strengthening it” (Perezgonzales, 2015)

Interpreting non-significance: Neyman-Pearson

Two hypotheses: H_M and H_A

Testing leads to accepting one of the hypotheses

Requires a priori calculation of power, expected minimum effect size

Sharply defined risk rates for false positives (Type I error, alpha) and false negatives (Type II error, beta)

No gradations of alpha: choose one beforehand (e.g. .01) and stick to it

With alpha, set up a critical value of a test for deciding between hypothesis (e.g., H_M: $M_1–M_2 = 0 \pm MES$, $\alpha = 0.05$, $CV_t = 2.38$)

P-values are proxies for critical values and have no evidential value

Interpreting non-significance: Neyman-Pearson

1. “If the observed result falls within the critical region, reject the main hypothesis and accept the alternative hypothesis.” (p < α)
2. “If the observed result falls outside the critical region and the test has good power, accept the main hypothesis.” (p > α)
3. “If the observed result falls outside the critical region and the test has low power, conclude nothing. (Ideally, you would not carry out research with low power).“ (Perezgonzales, 2015)

Interpreting non-significance: NHST

Null Hypothesis Significance Testing

Controversial amalgam of Fisher and Neyman-Pearson

Commonly taught to students and used in journals

Statistical significance (F) used for deciding between hypotheses (N-P)

$H_M = H_0$

H_A mostly as ’no H_0’

Sig = α, can be graded (* $p < .05$, ** $p < .01$, *** $p < .001$)

Non-significant results: (1) ignore and conclude nothing or (2) accept H_0

Interpreting non-significance: Bayes

P values: $P(D|H_0)$

Bayes: $P(H_0|D)$

Interpreting the probability of a hypothesis based on new data requires knowledge, beliefs or assumptions about the prior probability of the hypothesis

Assuming H_M and H_A are equally likely *a priori,* it is possible the main (or null) hypothesis becomes more probable when a non-significant result is observed

(https://www.r-bloggers.com/the-relation-between-p-values-and-the-probability-h0-is-true-is-not-weak-enough-to-ban-p-values/)

Testing whether there are no meaningful effects

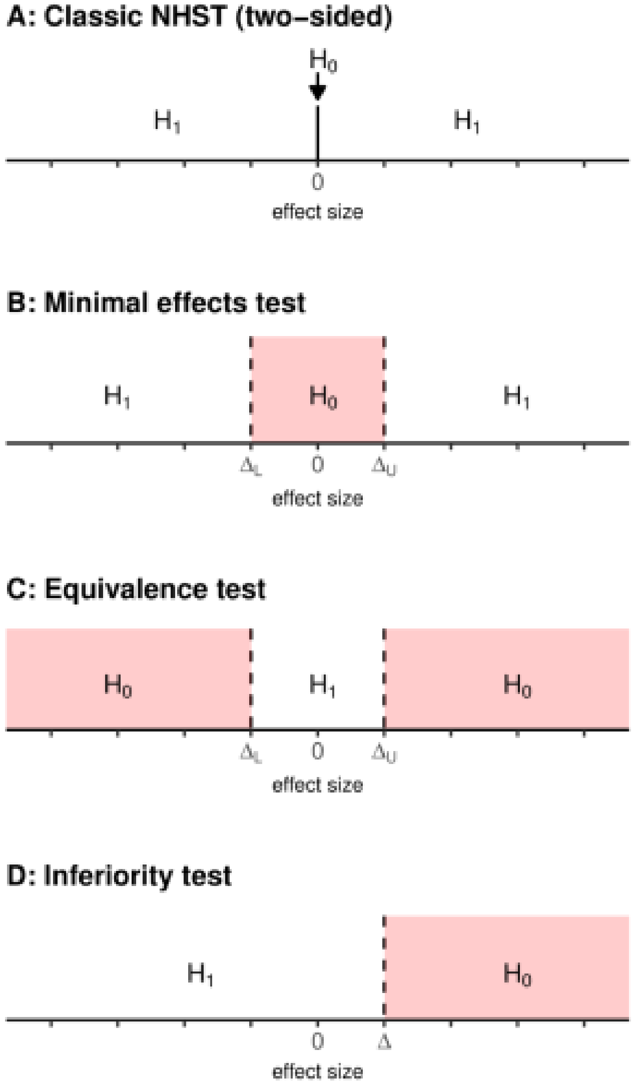


Figure 1: Illustration of null hypotheses (H_0) and alternative hypotheses (H_1) for different types of significance tests. **A)** NHST: Tests if the hypothesis (H_0) that an effect is equal to 0 can be rejected. **B)** Minimal effects test: Tests if the hypothesis (H_0) that an effect is larger than Δ_L *and* smaller than Δ_U can be rejected. **C)** Equivalence test: Tests if the hypothesis (H_0) that an effect is smaller than Δ_L *or* larger than Δ_U can be rejected. **D)** Inferiority test: Tests if the hypothesis (H_0) that an effect is larger than Δ can be rejected.

(Lakens & al., 2017; https://psyarxiv.com/v3zkt/)

Equivalence testing or TOST (Lakens & al., -17)

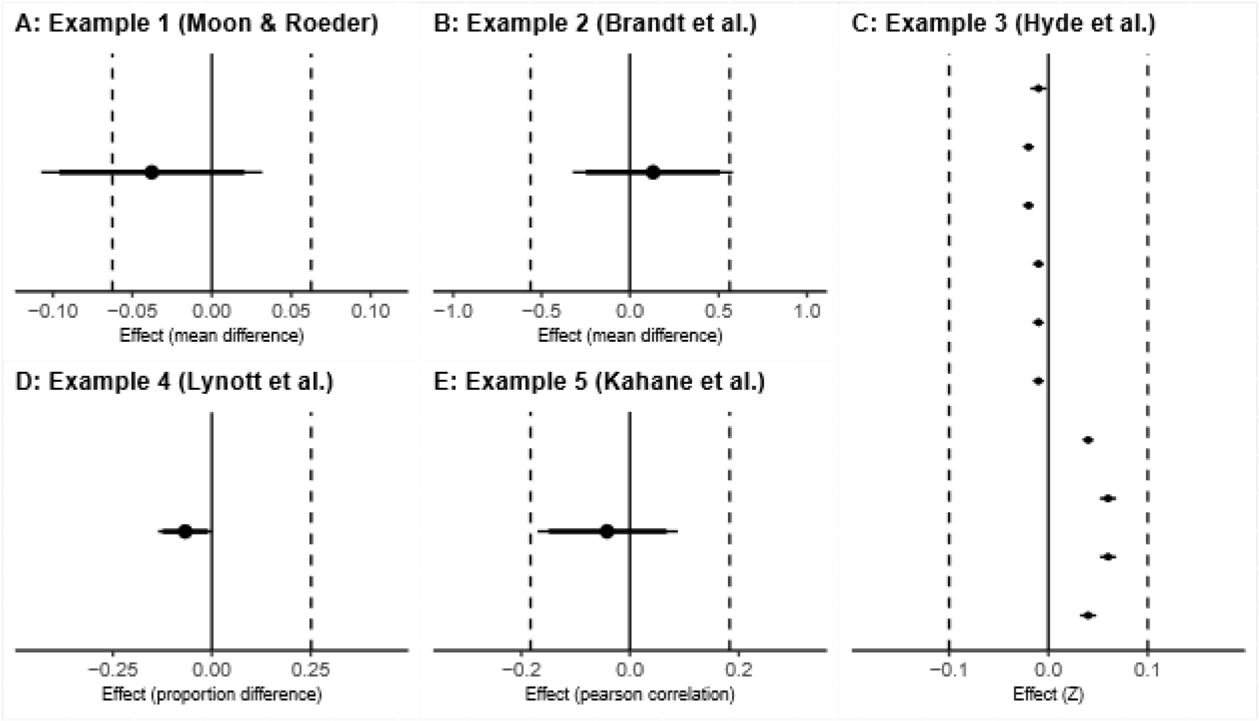


Figure 2: Example effects plotted with 90% TOST CIs (thick lines) and 95% NHST CIs (thin lines), the NHST null hypothesis (solid vertical line) and the equivalence bounds (dashed vertical lines) displayed. **A)** Example 1 - Mean Difference. **B)** Example 2 - Mean Difference. **C)** Example 3 - Meta-analytic Effect Size. **D)** Example 4 - Proportion Difference. **E)** Example 5 - Pearson Correlation.

Comparison

Approach	**Interpretation of *p* > .05**
Fisher	Non-significant results do not generally affect beliefs about null hypothesis
Neyman-Pearson	If the study power is acceptable, accept the main (null) hypothesis as true. Risk for a false negative is low enough.
NHST	Ignore or accept the null hypothesis
Bayes	Almost all data affect the probability estimates of hypotheses, including *p* > .05
Equivalence testing	Support for the hypothesis of no effect requires that the result is a precise enough zero

Interpreting replication failure

”It is too easy to conclude that successful replication means that the theoretical understanding of the original finding is correct.” (OSC, 2015)

”It is also too easy to conclude that conclude that a failure to replicate a result means that the original evidence was a false positive.” (OSC, 2015)

” After this intensive effort to reproduce a sample of published psychological findings, how many of the effects have we established are true? Zero. And how many of the effects have we established are false? Zero. Is this a limitation of the project design? No.” (OSC, 2015)

However, results consistent with

low power + publication bias =

upwardly biased effect sizes + irreproducible reseach

Interpreting replication failure

Maximum reproducibility of original results is not always important (OSC, 2015)

Exploratory and daring but non-replicable research is not bad

Exploratory research should not be presented as confirmatory regardless of its replicability

Polls

PollEv.com/esapalosaari182

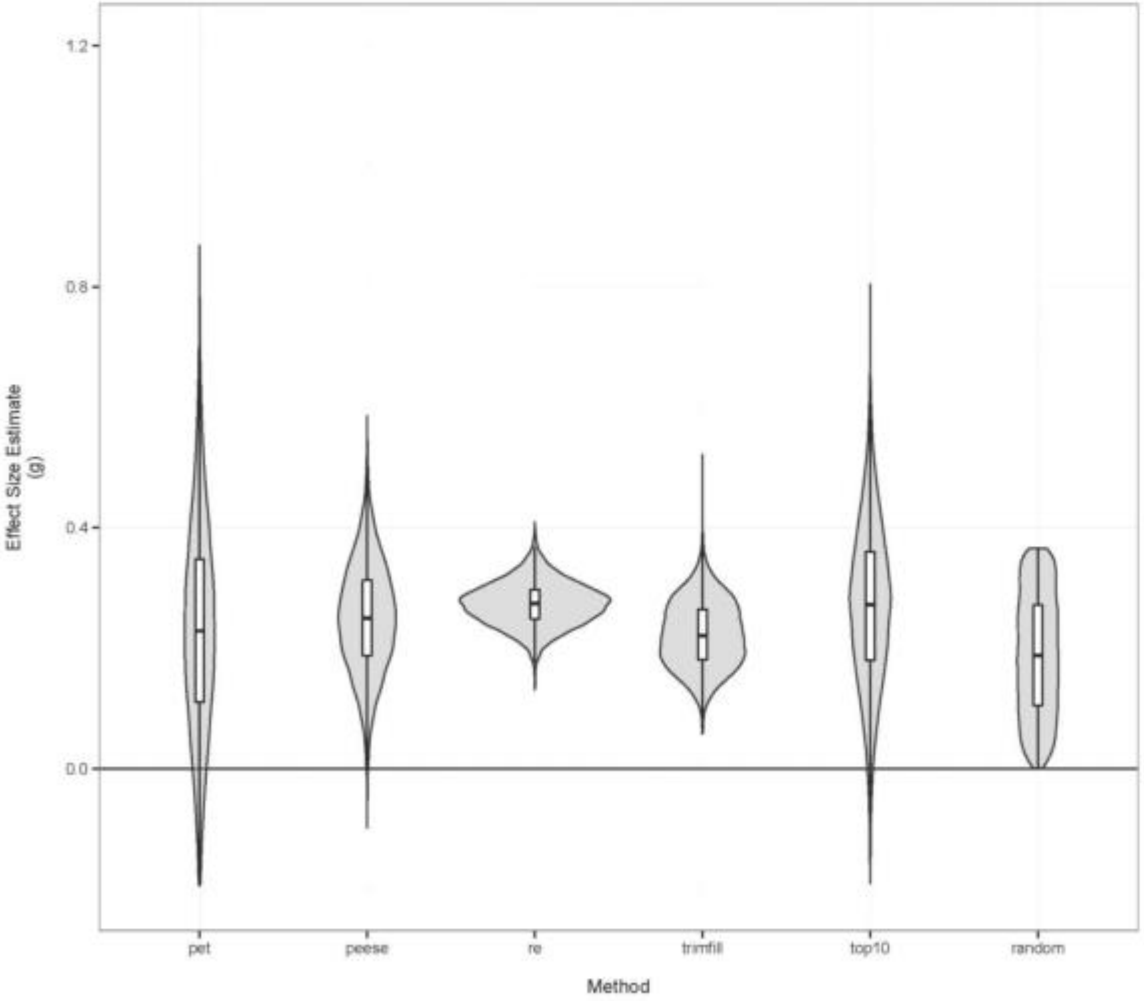
Improving statistical inferences:
P-curve analysis and preregistration

P-curve analysis: Motivation

How to determine whether a ’contaminated’ literature has evidential value?

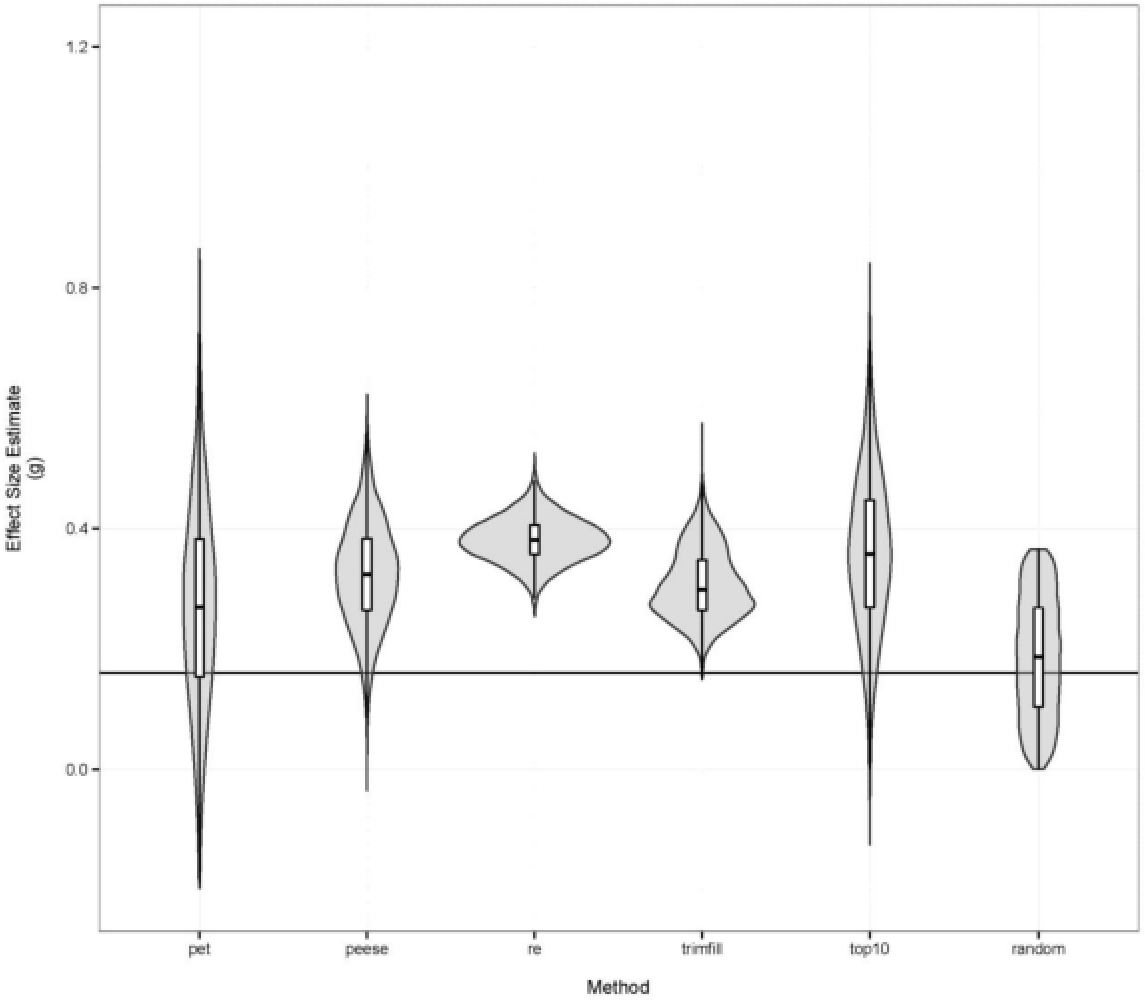
Meta-analyses are unable to correct for publication bias and flexibility in data analysis (Inzlicht, Gervais & Berkman, 2015)

Figure 1a. Violin plots depicting the effectiveness of various meta-analytic techniques to estimate the size of an effect, when the true effect (solid horizontal line) is nil (g=0).



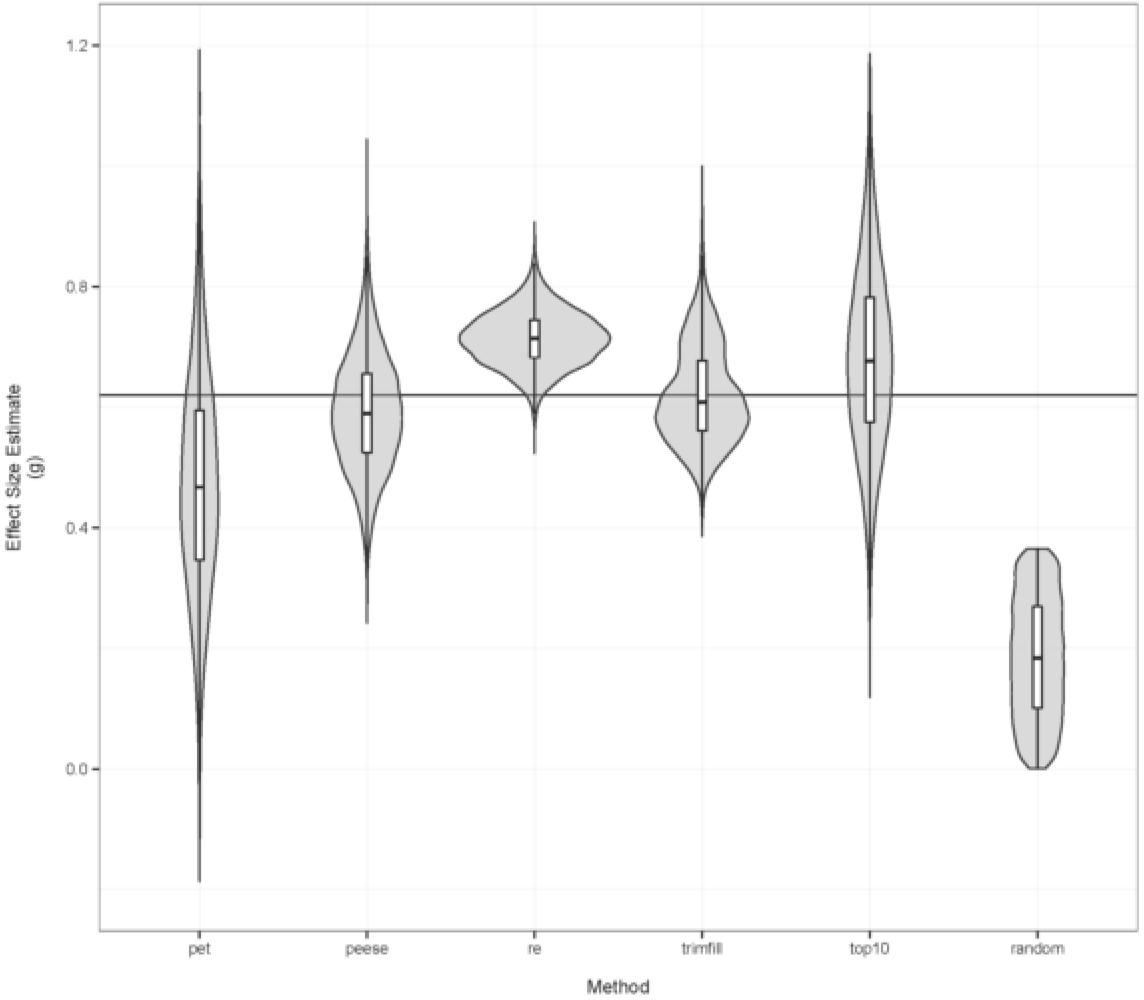
Note: The width of each plot corresponds to the frequency of estimates at that level. PET = Precision Effect Test. PEESE =Precision Effect Estimation with Standard Error test. RE = Random Effect meta-analysis. Trimfill = Trim and Fill. Top10 = Top 10 estimator. Random = Random correction.

Figure 1b. Violin plots depicting the effectiveness of various meta-analytic techniques to estimate the size of an effect, when the true effect (solid horizontal line) is small (g=.16).



Note: The width of each plot corresponds to the frequency of estimates at that level. PET = Precision Effect Test. PEESE =Precision Effect Estimation with Standard Error test. RE = Random Effect meta-analysis. Trimfill = Trim and Fill. Top10 = Top 10 estimator. Random = Random correction.

Figure 1d. Violin plots depicting the effectiveness of various meta-analytic techniques to estimate the size of an effect, when the true effect (solid horizontal line) is large (g=.62).



P-curve analysis: Idea

Maybe analysing the distribution of *p*-values under .05 can give us information about the true true effect (Simonsohn, Nelson, & Simmons, 2014)?

Only true effects are expected to generate right-skewed *p*-curves (more .01s than .04s)

P-curve analysis: Simulated *p*-hacking

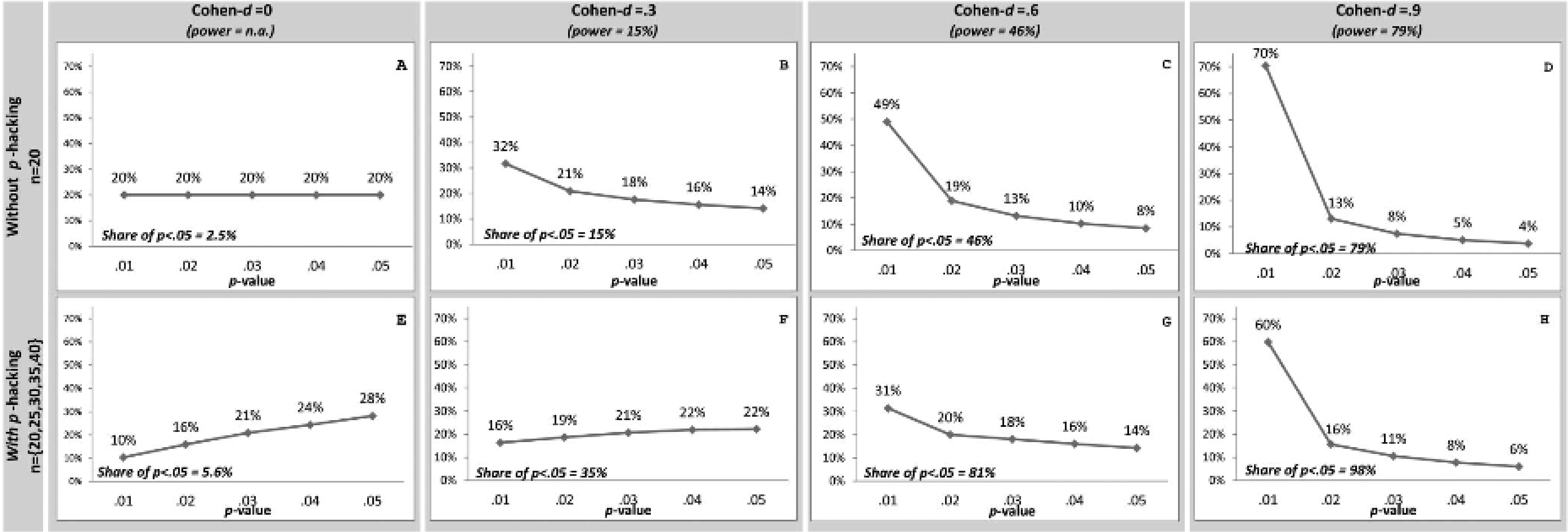
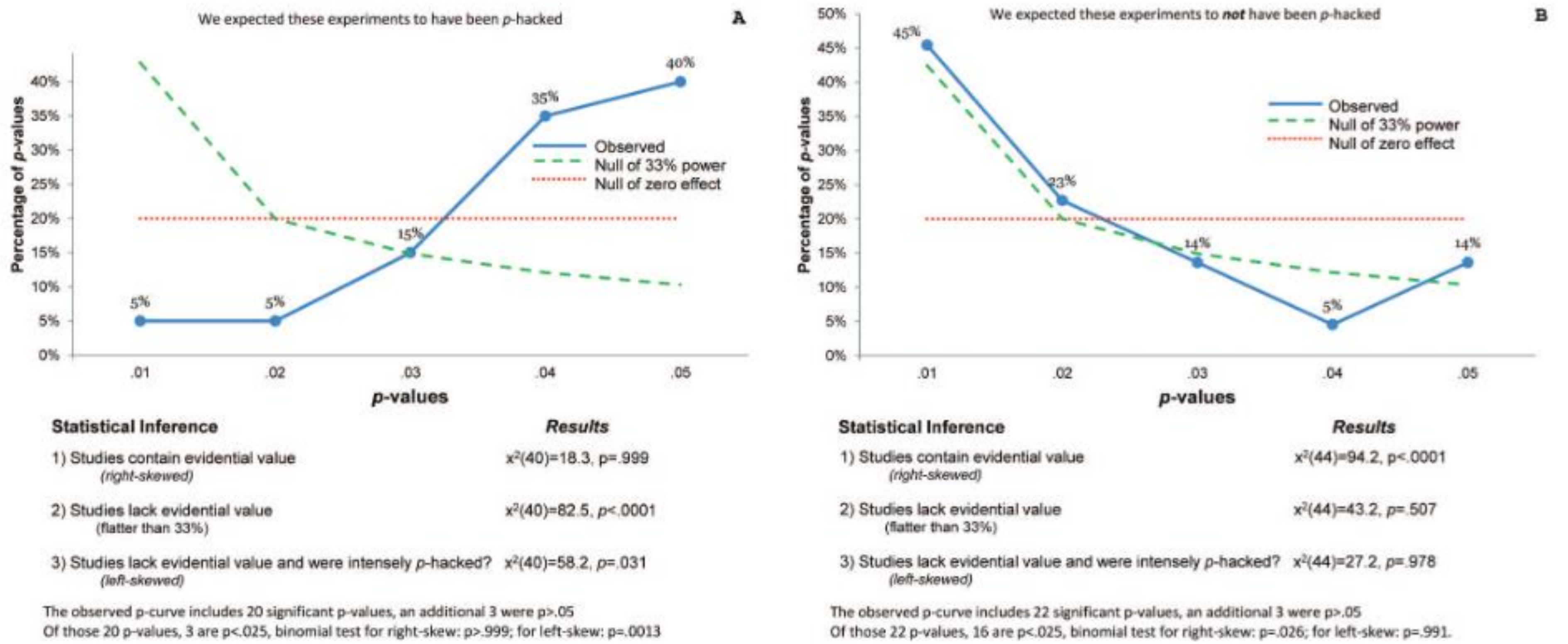


Figure 1. *P*-curves for different true effect sizes in the presence and absence of *p*-hacking. Graphs depict expected *p*-curves for difference-of-means *t* tests for samples from populations with means differing by *d* standard deviations. A–D: These graphs are products of the central and noncentral *t* distribution (see Supplemental Material 1). E–H: These graphs are products of 400,000 simulations of two samples with 20 normally distributed observations. For 1E–1H, if the difference was not significant, five additional, independent observations were added to each sample, up to a maximum of 40 observations. Share of $p < .05$ indicates the share of all studies producing a statistically significant effect using a two-tailed test for a directional prediction (hence 2.5% under the null).

P-curve analysis: Real data



Statistical Inference	***Results***
1) Studies contain evidential value *(right-skewed)*	$x^2(40)$=18.3, p=.999
2) Studies lack evidential value (flatter than 33%)	$x^2(40)$=82.5, *p*<.0001
3) Studies lack evidential value and were intensely *p*-hacked? *(left-skewed)*	$x^2(40)$=58.2, *p*=.031

The observed p-curve includes 20 significant p-values, an additional 3 were p>.05
Of those 20 p-values, 3 are p<.025, binomial test for right-skew: p>.999; for left-skew: p=.0013

Statistical Inference	***Results***
1) Studies contain evidential value *(right-skewed)*	$x^2(44)$=94.2, p<.0001
2) Studies lack evidential value (flatter than 33%)	$x^2(44)$=43.2, *p*=.507
3) Studies lack evidential value and were intensely *p*-hacked? *(left-skewed)*	$x^2(44)$=27.2, *p*=.978

The observed p-curve includes 22 significant p-values, an additional 3 were p>.05
Of those 22 p-values, 16 are p<.025, binomial test for right-skew: p=.026; for left-skew: p=.991.

P-curve analysis: Tests

How to test if *p*-curve is significantly right- (or left-) skewed?

One method is to divide the *p* values as high ($p > .025$) or low ($p < .025$) and do a binomial test

Another method Simonshon & al. propose is to calculate the probability of observing a *p* value at least as extreme if the null hypothesis were true, a *p* value of *p* value

The *pp* values are aggregated giving a chi^2 test for skew

P-curve analysis: Simulated tests

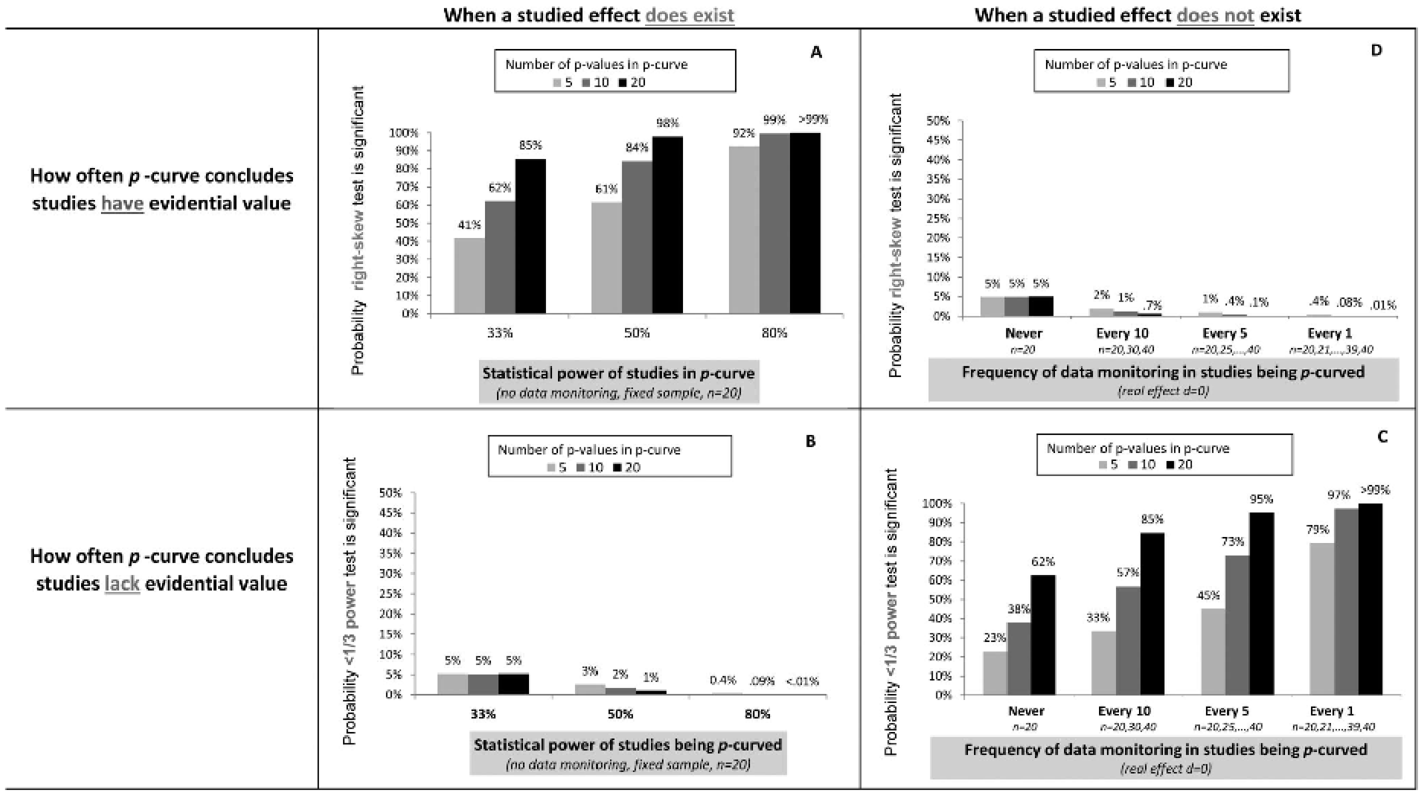


Figure 6. This figure shows how often *p*-curve correctly and incorrectly diagnoses evidential value. The bars indicate how often *p*-curve would lead one to conclude that a set of findings contains evidential value (a significant right-skew; A & D) or does not contain evidential value (powered significantly below 33%; B & C). Results are based on 100,000 simulated *p*-curves. For A and B, the simulated *p*-curves are derived from *p*-values drawn at random from noncentral distributions. For C and D, *p*-curves are derived from collecting *p* values from simulations of *p*-hacked studies. The *p*-hacking is simulated the same way as in Figure 1.

P-curve analysis: Assumptions

Included *p* values must meet three criteria:

1. Test the hypothesis of interest (not unrelated studies)
2. Have a uniform distribution under the null (no discrete variables)
3. Be statistically independent from other selected *p*-values

P-curve analysis: Steps to do it

1. Set a rule for selecting studies in advance

   ”All studies published in 2009 with wine as a manipulation and simulated driving behavior as a dependent variable.”

2. Create a *P*-curve Disclosure Table to select the results to analyze
   1. Identify researchers’ stated hypothesis and study design quoting from paper
   2. Identify the statistcial reult testing stated hypothesis using Table 3 in the Guide (http://p-curve.com/guide.pdf)
   3. Report the statistical results of interest quoting from paper
   4. Recompute the precice *p*-value(s) based on reported test statistics
   5. Report robustness results (with and without ambiguous inclusions)

3. Feed key results to *p*-curve app (www.p-curve.com/app4)

4. Copy-paste app’s output onto your paper

P-curve analysis: Criticism

- Problems claimed in a presentation a couple of weeks ago (http://richarddmorey.org/content/Psynom17/pcurve/#/)
- Errors in constructing the tests
  - Over-sensitivity to values near alpha (.05)
- Lack of justificaction for meta-analytic grouping
  - How to solve debates over ’proper’ groupings?

Preregistration: Idea

- A time-stamped, read-only version of a research plan created before the study
- Increasing the credibility of research by specifying in advance how data will be analyzed
- A potential solution to
    - Selective analysis (the garden of forking paths, *p*-hacking)
    - Selective reporting (the file drawer, publication bias)
    - Hypothesizing After Results are Known (HARKing)

Preregistration: Idea

- Preregistration makes the distinction between

  hypothesis testing (confirmatory) and

  hypothesis generating (exploratory) research clear
- Backing to claims ”as predicted … ” or ”contrary to expectations …”
- There’s a difference between predicting yesterday’s and tomorrow’s stock market

Preregistration: Examples

- https://aspredicted.org/
  - Answer 9 questions
  - Stays private until an author act to make it public
  - Authors may share anonymous .pdf with reviewers
- https://osf.io/
- https://osf.io/prereg/
  - Need to create an account and a project
  - Answer much more detailed questions
    - Sampling plans
    - Variables
    - Design plan
    - Analysis plan
    - Scripts

Preregistration: Criticism

- Too restrictive?
  - Do a preregistered confirmatory analysis to a part of the study, explore the rest
  - Or: split the dataset to a an exploratory (training set) and confirmatory (validation set) part
- Doesn’t really prevent cheating
  - It is possible to cheat: create multiple preregistrations and publish only some
  - Biomedical registries are one piece of evidence about publication bias
  - Perhaps the main potentially beneficial effect could be through keeping researchers honest towards themselves?

Other solutions

- Registered Reports
  - Similar to preregistrations
  - Additionally: improvement of study plans via peer-review
  - Journals promise to publish regardless of the significance of results
- …?

References

Camerer, C. F. & al. (2016). Evaluating replicability of laboratory experiments in economics. Science, aaf0918. DOI: 10.1126/science.aaf0918

Chan, A.-W. & Altman, D. G. (2005). Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. BMJ, 330(7494), 753. 10.1136/bmj.38356.424606.8F

Decullier, E. & Chapuis, F. (2005). Fate of biomedical research protocols and publication bias in France: retrospective cohort study. BMJ, 331, 19. https://doi.org/10.1136/bmj.38488.385995.8F

Dickersin, K. & Min, Y. (1993). NIH clinical trials and publication bias. Online J Curr Clin Trials, 50.

Gelman, A. & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem even when there is no ”fishing expedition” or ”*p*-hacking” and the research hypothesis was posited ahead of time. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Inzlicht, Michael and Gervais, Will and Berkman, Elliot, Bias-Correction Techniques Alone Cannot Determine Whether Ego Depletion is Different from Zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015 (September 11, 2015). Available at SSRN: https://ssrn.com/abstract=2659409 or http://dx.doi.org/10.2139/ssrn.2659409

Lakens, D., Scheel, A. & Isager, P. (2017). Equivalence testing for psychological research: A tutorial. DOI: 10.17605/OSF.IO/V3ZKT

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? American Psychologist, 70, 487-498.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615-631.

Perezgonzales, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6:* 223.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251).

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. Journal of Experimental Psychology: General, 143(2), 534-547. http://dx.doi.org/10.1037/a0033242

Song, F., Parekh-Bhurke, S, … , Harvey, I. (2009). Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. BMC Medical Research Methodology, 9, 79. https://doi.org/10.1186/1471-2288-9-79