Practical Skills for Navigating the Crisis

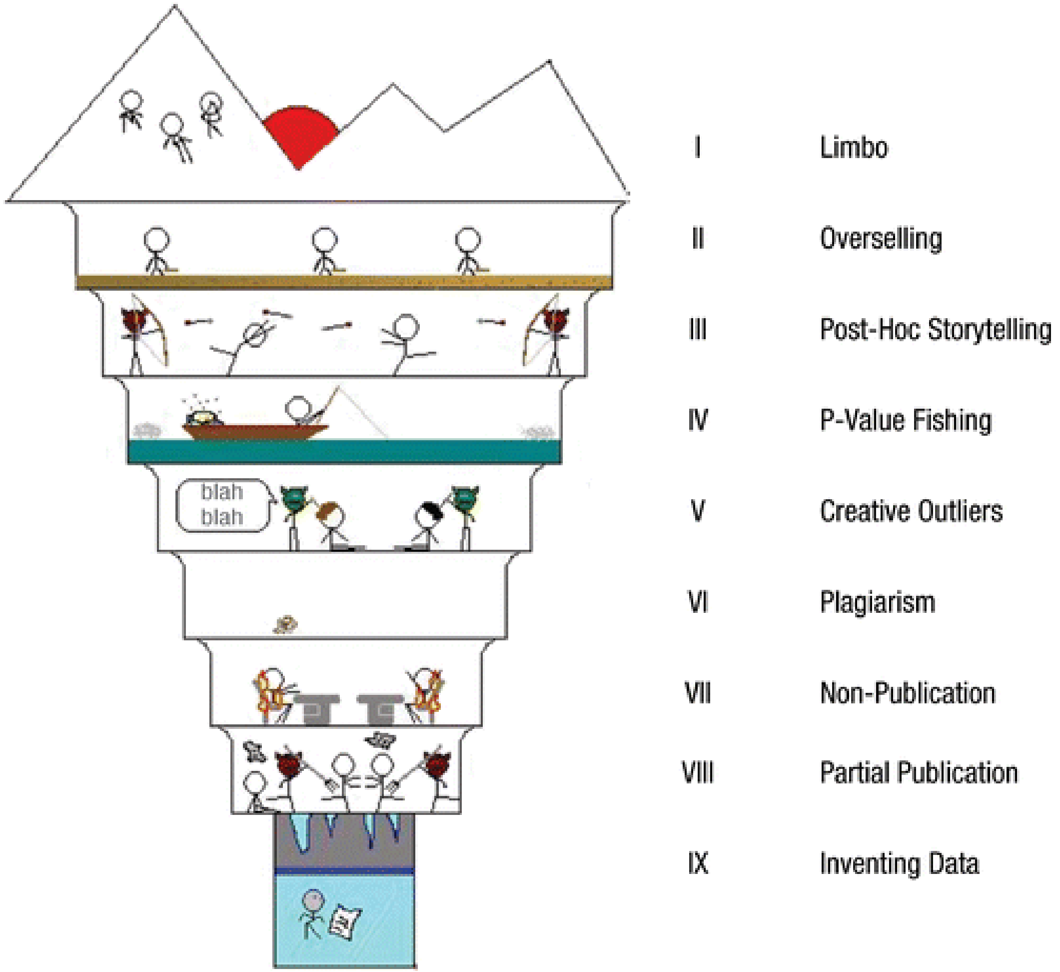
Detecting low credibility research and doing high credibility research

Noah Jones

1/13/2021

MAS.S73

Nine circles of scientific hell



Spectrum of impact to scientific career

Neuroskeptic et al. 2012

Factors leading to replication failures

I. Uncommon

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Factors leading to replication failures

I. Uncommon

II. Common, but not primary culprit

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Factors leading to replication failures

I. Uncommon

II. Common, but not primary culprit

III. Very common

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Factors leading to replication failures

I. Uncommon

Fraud

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Factors leading to replication failures

I. Uncommon

Fraud

II. Common, but not primary culprit

File-drawering failed studies

Innocent Errors

Insufficient power

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Factors leading to replication failures

I. Uncommon

Fraud

II. Common, but not primary culprit

File-drawering failed studies

Innocent Errors

Insufficient power

III. Very common

P-hacking

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Methods to tackle each potential problem

I. Uncommon

Cataloguing publications that need to be retracted

Disclosure of all data and materials

Fraud detection flags

II. Common, but not primary culprit

III. Very common

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Methods to tackle each potential problem

I. Uncommon

Cataloguing publications that need to be retracted

Disclosure of all data and materials

Fraud detection flags

II. Common, but not primary culprit

Pre-registration & Journal acceptance

Statistical checks

Power analysis

III. Very common

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Methods to tackle each potential problem

I. Uncommon

Cataloguing publications that need to be retracted

Disclosure of all data and materials

Fraud detection flags

II. Common, but not primary culprit

Pre-registration & Journal acceptance

Statistical checks

Power analysis

III. Very common

Meta-Statistics

(Pre-registration/Registered reports)

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Publication Transparency Databases

- Publication Meta Data
  - Pre-registration
  - Data and Materials

Publication Transparency Databases

- Publication Meta Data
  - Pre-registration
  - Data and Materials
- Retractions

Publication Transparency Databases

- Publication Meta Data
  - Pre-registration
  - Data and Materials
- Retractions
- Replications

Publication Transparency Databases

- Publication Meta Data
  - Pre-registration
  - Data and Materials
- Retractions
- Replications
- Commentary

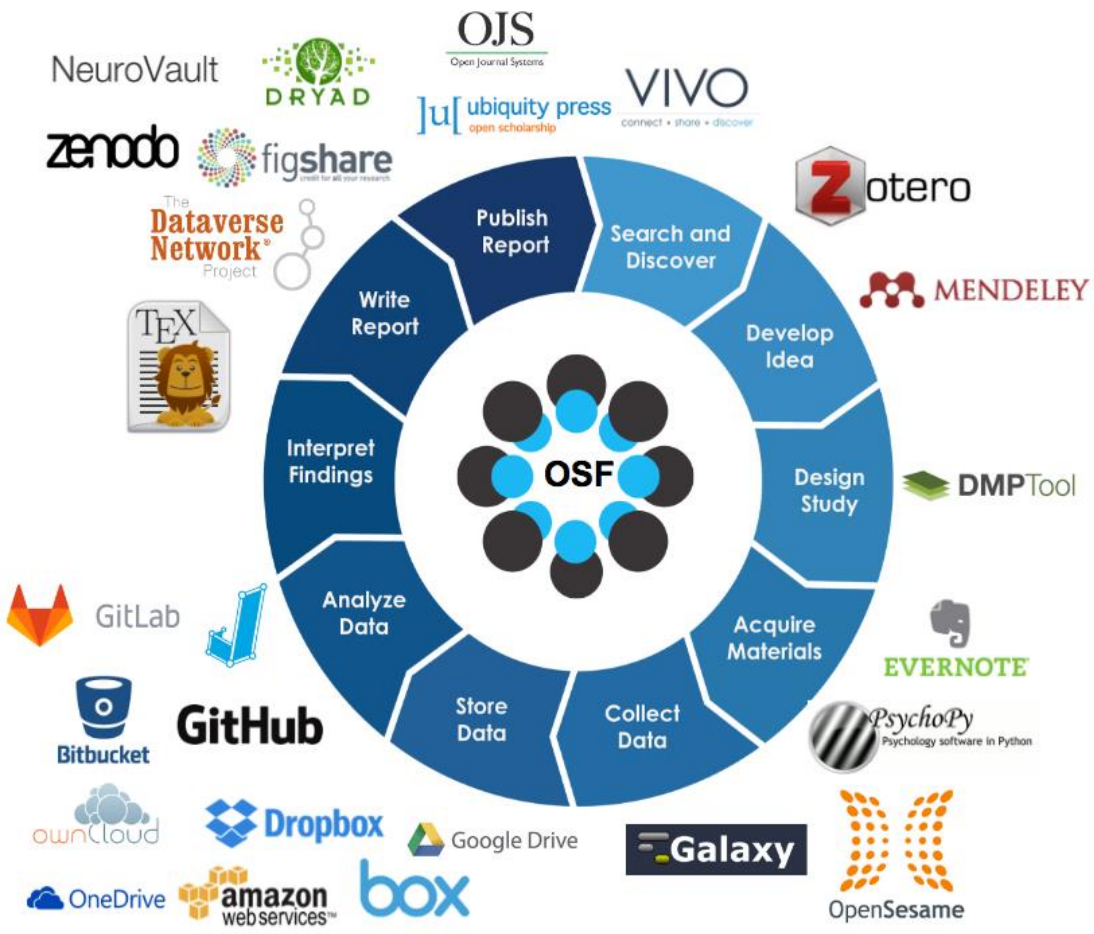
Meta Data: Open Science Foundation

http://osf.io

Meta Data: Open Science Foundation

Meta Data: Open Science Foundation

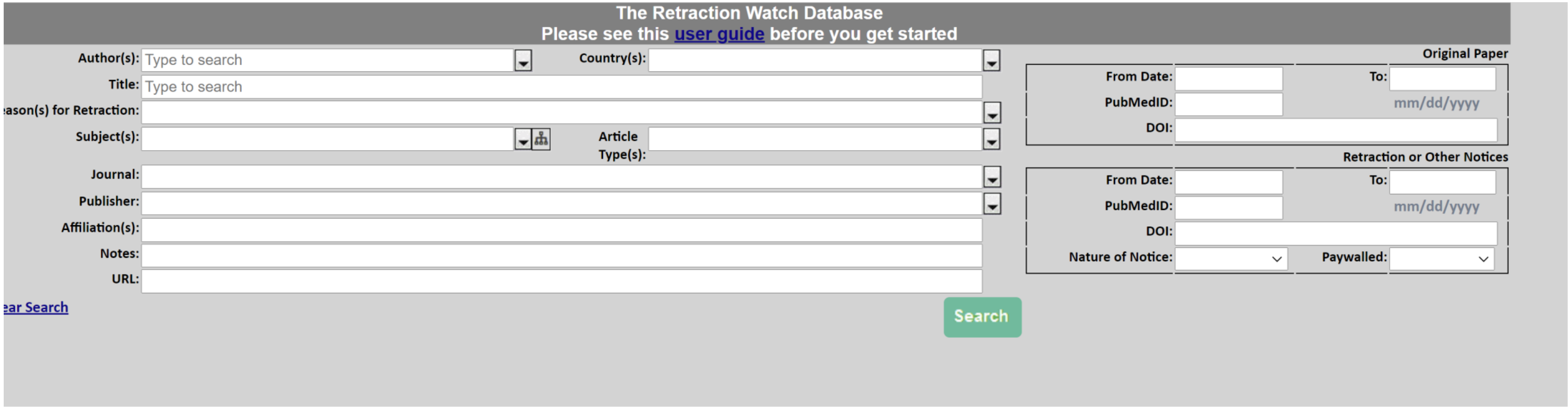
Integrations



Retractions: Retraction Watch

http://www.retractionwatch.com

30K Retractions in Database



Replications:

Forrt

Framework for
Open and
Reproducible
Research
Training

FORRT

Replications:

Forrt

Data Replicada

DATA COLADA

Thinking about evidence, and vice versa

Replications:

Forrt

Data Replicada

Multi-Lab Groups

Registered Replication Report

Many Labs 2: Investigating Variation in Replicability Across Samples and Settings

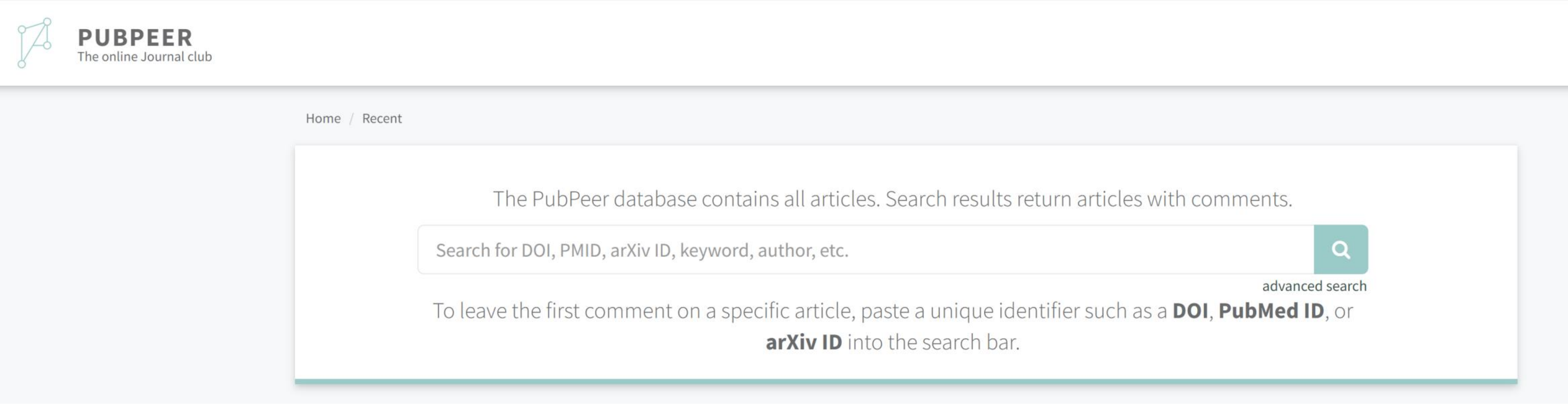
Psychological Science Accelerator

A Distributed Laboratory Network

study swap

Commentary: PubPeer

http://Pubpeer.com



Commentary: PubPeer

Tonic inhibition enhances fidelity of sensory information transmission in the cerebellar cortex

Journal of Neuroscience (2012) - 6 Comments

pubmed: 22875944 doi: 10.1523/jneurosci.0460-12.2012 issn: 0270-6474 issn: 1529-2401

Ian Duguid, Tiago Branco, Michael London, Paul Chadderton, Michael Häusser

#1 ***Peer 1*** commented December 2012

It is surprising to see in figure 2 that sensory input provides neither feed-forward nor feedback inhibition onto granule cells. Does this suggest that the Golgi cell's role in the circuit is only to set the amplitude of tonic inhibition?

report permalink Reply

Commentary: Curate Science

http://curatescience.org

MISSION

Accelerate science by developing the best *transparency and credibility curation tools* for all research stakeholders.

VISION

Create an *accountable* research world brimming with *transparent* and *credible* evidence.

Every year, millions of people suffer and/or die from serious conditions like cancer, Alzheimer's, heart disease, anxiety/mood disorders, and suicide. To make progress on these and other problems, funded scientific research must be, at minimum, **transparent and credible** (credible research is transparent evidence that survives scrutiny from peers). Transparent and credible evidence can then be built upon, which allows ever more precise theories/hypotheses to be tested (solid cumulative knowledge cannot be built on quicksand). Sadly, there is a growing body of compelling evidence that a great deal of current academic research (if not the majority:1, 2) is neither minimally transparent nor credible (1, 2, 3, 4, 5, 6, 7, 8, 9,10, 11, 12, 13, 14, 15, 16). Worse, there's no systematic way to differentiate credible evidence from untrustworthy evidence.

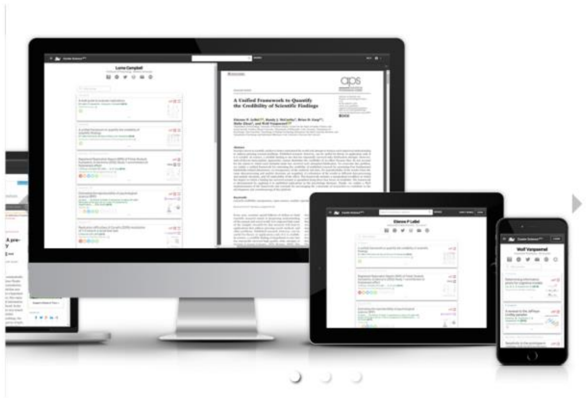
Curate Science is an integrated system and curation platform to verify that research is **transparent and credible** (for a visual overview, see hyperlinked diagram). It will allow researchers, journals, universities, funders, teachers, journalists, and the general public to ensure:

1. Transparency: Ensure research meets minimum transparency standards appropriate to the article type and employed methodologies.
2. Credibility: Ensure follow-up scrutiny is linked to its parent paper, including critical commentaries, reproducibility/robustness re-analyses, and new sample replications.

This will ensure that researchers, journals, universities, and funders are **accountable** to the people they serve. A unified platform to differentiate *credible evidence* (from untrustworthy evidence) will substantially accelerate the development of cumulative scientific knowledge and applied innovations across the natural and social sciences. The implications for human welfare are large.

Commentary: Curate Science

FOR AUTHORS



Organize your publications on your own Curate Scholar author page to make your science deliciously user-friendly, ultimately accessible, and beautiful on all devices (example author pages: 1, 2, 3).

View full-text PDF and HTML versions of your articles directly within your author page.

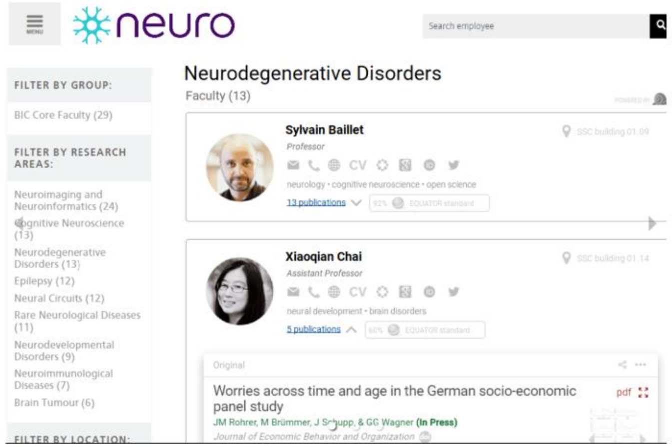
Expose key figures in your publication list so your readers can jump directly into your research via a delightful touch-enabled media viewer.

Curate links to associated content to save your reader and yourself time (e.g., URLs to open data, talks/videos).

CREATE AUTHOR PAGE

Commentary: Curate Science

FOR UNIVERSITIES



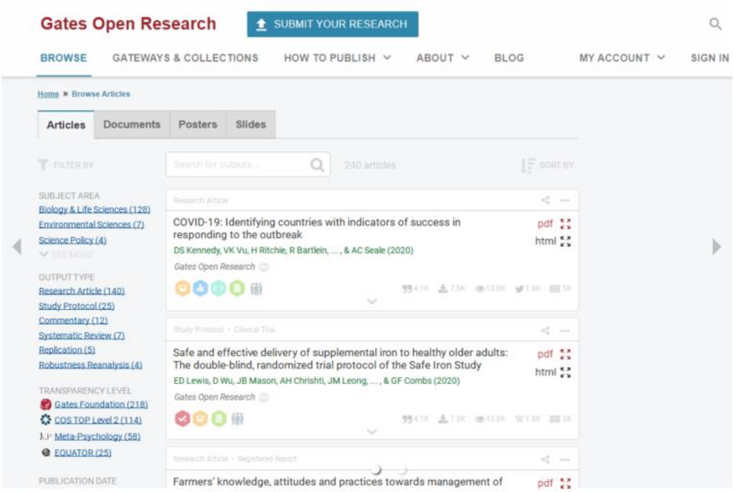
Make your researchers' publications easy to access, interactive, and deliciously user-friendly to consume on your university's departmental pages.

Track the open science practices of your researchers, and monitor your progress in achieving transparency targets prioritized by your institution (see interactive prototype).

University departments can then be ranked by their transparency track record, which graduate students and job candidates can use to inform their decisions at what university to work.

Commentary: Curate Science

FOR FUNDERS



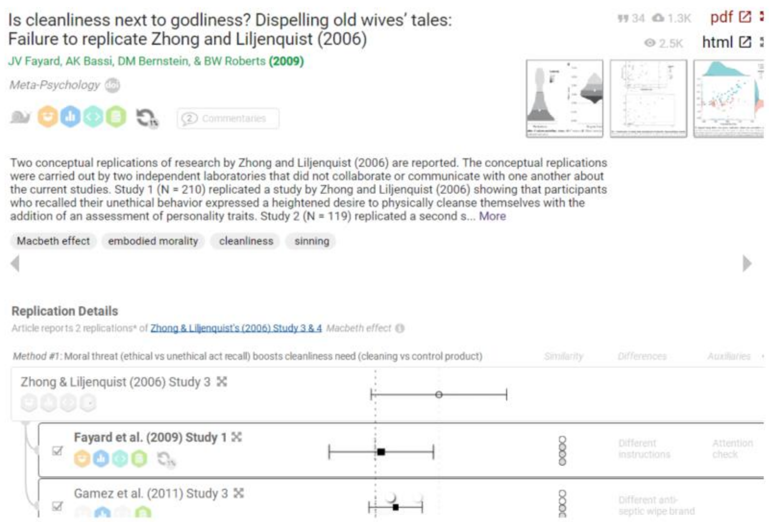
Make your grantees' research outputs easy to access, interactive, and deliciously user-friendly to consume for all research stakeholders (e.g., policy analysts, innovators, citizens, etc.).

Track the transparency of the research you fund, and monitor your progress in requiring higher levels of research transparency (see interactive prototype).

Monitor your progress in funding a larger proportion of studies that report independent replications and reproducibility re-analyses.

Commentary: Curate Science

FOR REPLICATORS



Link your replication to the original study to increase its visibility, discoverability, and impact, accelerating scientific self-correction.

Curate replication metadata on its own article page and easily share it.

Create collections of replications across different methods of testing an effect, and meta-analyze and track replication evidence (coming soon).

Fraudulent and inconsistent Data

Numerical Tests

- GRIM Test

The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology

Nicholas J L Brown[1], James A J Heathers[2]

Fraudulent and inconsistent Data

- Numerical Tests
  - GRIM Test
- Image Manipulation
  - Adobe Bridge and ImageJ

Fraudulent and inconsistent Data

- Numerical Tests
  - GRIM Test
- Image Manipulation
  - Adobe Bridge and ImageJ
- Stat checking
  - Statcheck.io

Methods to tackle each potential problem

I. Uncommon

Cataloguing publications that need to be retracted

Disclosure of all data and materials

Fraud detection flags

II. Common, but not primary culprit

Pre-registration & Journal acceptance

Statistical checks

Power analysis

III. Very common

Meta-Statistics

(Pre-registration/Registered reports)

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Preregistration

Separates Hypothesis Generating
(**Exploratory Research**)

Hypothesis-testing (**Confirmatory Research**)

What is a preregistration

Research plan

- Time-stamped
- Immutable or read-only
- Created before the study
- Submitted to public registry

Benefits of Preregistration

Can protect against natural biases and selective reporting

Great tool for communicating work with others

More robust planning

Helpful reminder of what you plan

What does it contain

Study Plan

- Hypothesis
- Data collection procedures
- Manipulated and measured variables

Analysis Plan

- Statistical model
- Inference criteria

Examples of preregistration

https://osf.io/h9k8n/

OSF Preregistration Templates

https://osf.io/zab38/wiki/home/

Problems with preregistration

[How to Crack Pre-registration: Toward Transparent and Open Science (Yamada et al. 2018)]

Yamada argues to deal with these challenges we should have **journals** for experimental or confirmatory research **and** theoretical or exploratory research

Registered Reports

A **stronger** preregistration

"Registered Reports eliminates the bias against negative results in publishing because the results are not known at the time of review."

-- Daniel Simons, Professor at University of Illinois, Urbana-Champaign, co-editor of Registered Replication Reports at Perspectives on Psychological Science, and incoming chief editor of Advances in Methods and Practices in Psychological Science

"Because the study is accepted in advance, the incentives for authors change from producing the most beautiful story to the most accurate one."

--Chris Chambers, Professor at Cardiff University, Section Editor for Registered Reports at Cortex, European Journal of Neuroscience and Royal Society Open Science, Chair of the Registered Reports Committee supported by the Center for Open Science

Resources

https://aspredicted.org/

https://osf.io/prereg/

https://www.cos.io/blog/preregistration-plan-not-prison

https://cos.io/prereg

https://www.cos.io/initiatives/registered-reports (Database of journals accepting registered reports)

Transparent and Reproducible Social Science Research: How to Do Open Science (Christensen et al.)

Power Analysis:
Possible conclusions from a test

<table>
<tr><td colspan="2"></td><td colspan="2">Null Hypothesis (H_0) is:</td></tr>
<tr><td colspan="2"></td><td>True</td><td>False</td></tr>
<tr><td rowspan="2">Judgment of Null (Statistical Result)</td><td>Reject H_0
($p < .05$)</td><td></td><td></td></tr>
<tr><td>Fail to reject H_0
($p > .05$)</td><td></td><td></td></tr>
</table>

Citation of material: OSF

Possible conclusions from a test

		Null Hypothesis (H_0) is:	
		True	False
Judgment of Null (Statistical Result)	Reject H_0 ($p < .05$)	Type I Error False Positive α	
	Fail to reject H_0 ($p > .05$)	Correct Inference True Negative	

Possible conclusions from a test

		Null Hypothesis (H_0) is:	
		True	False
Judgment of Null (Statistical Result)	Reject H_0 ($p < .05$)	Type I Error False Positive α	Correct Inference True Positive $(1 - \beta)$
	Fail to reject H_0 ($p > .05$)	Correct Inference True Negative	Type II Error False Negative β

Possible conclusions from a test

Power

		Null Hypothesis (H_0) is:	
		True	False
Judgment of Null (Statistical Result)	Reject H_0 ($p < .05$)	Type I Error False Positive α	Correct Inference True Positive $(1 - \beta)$
	Fail to reject H_0 ($p > .05$)	Correct Inference True Negative	Type II Error False Negative β

What is Power

Probability to reject the null hypothesis (H_0) is False given that it is False

80% Power means have **an 80% chance of getting significant result** when effect is true

Based on **effect size, sample size and alpha level**

Why is Power important?: Problems with Low Power

Increased likelihood of **false negative**

Inflated effect size when significance is there

Lower positive predictive value (true positives)

False Negatives

The lower the power of your study, the more likely you'll find a false negative

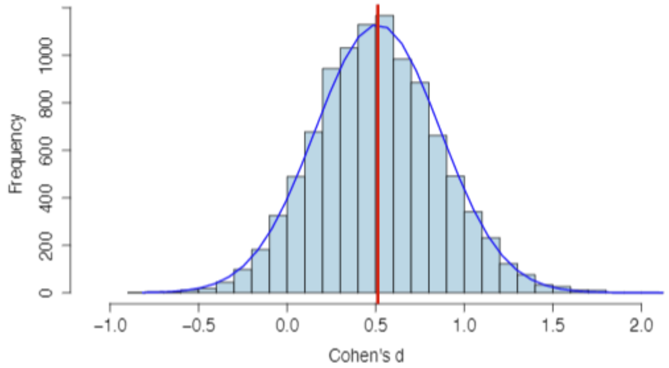
E.X not finding an average differences in height between men and women

Inflated Effect Size

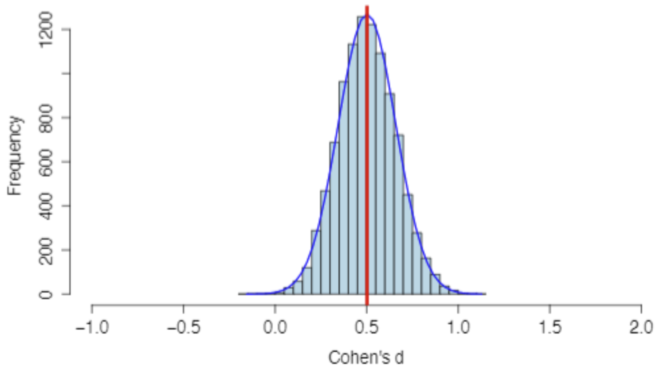
Samples drawn from population given effect size is distributed around true effect size

Power of studies does not affect distribution mean, but the shape and areas of significance in distribution

Distribution Shape

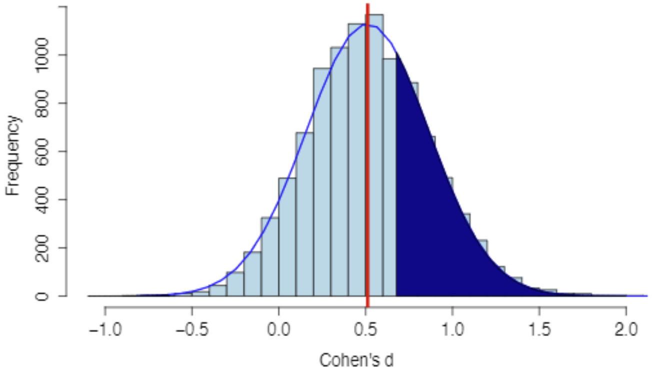


30% Power

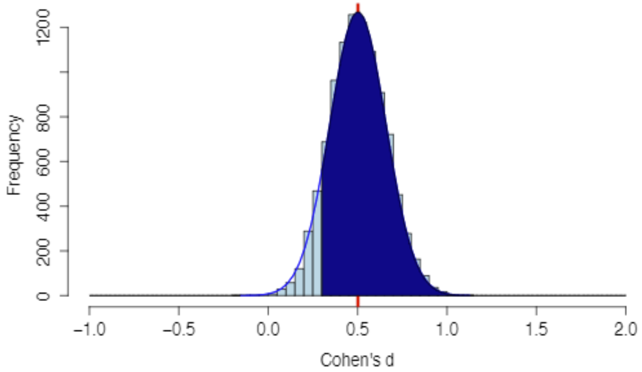


90% Power

Significant Effect Sizes



30% Power



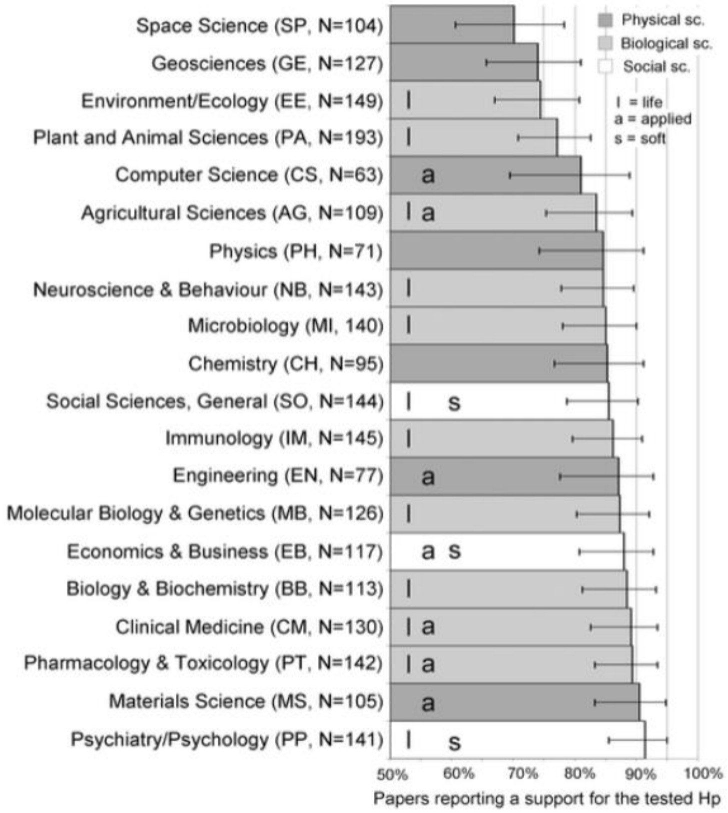
90% Power

Inflated Effect Sizes

As studies become more underpowered, only tails of distribution will reach statistical significance

- Leads to extreme inflation as power decreases

Inflated Effect Sizes



Fanelli D (2010)

We can overestimate the effectiveness of our treatments

It is difficult to properly power future studies based on past research (true power of a study using effect size from previous is likely lower than power analysis would suggest)

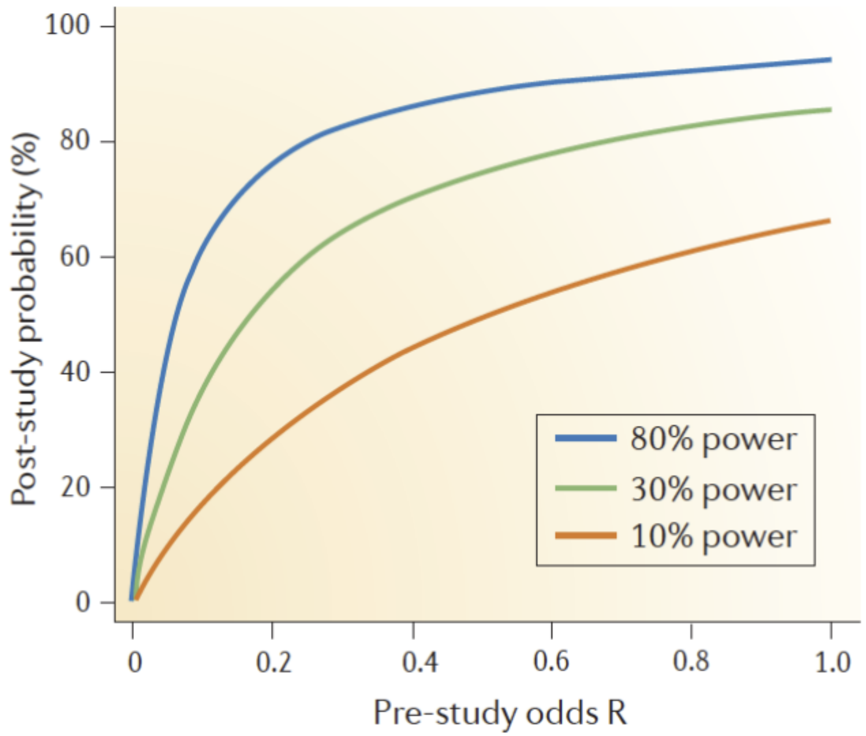
Positive Predictive Value

Probability that a positive result represents a true positive

- Effect is real in the population

$$PPV = \frac{(1-\beta) * OR}{[(1-\beta) * OR] + \alpha}$$

- OR: Odds that our hypothesis is true
- $(1 - \beta)$: Power
- A: Alphas level



Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson, & Munafo (2011)

Intro to Power Analysis

- Specify alpha level and power level
  - Usually set it to 0.05 and power to 0.80

Intro to Power Analysis

- Specify alpha level and power level
  - Usually set it to 0.05 and power to 0.80
- Get mean test scores between two groups

Intro to Power Analysis

- Specify alpha level and power level
  - Usually set it to 0.05 and power to 0.80
- Get mean test scores between two groups
- Compute expected effect size (Cohen's D or R)
  - Get N Values of two samples
  - Get standard deviations of scores
  - Means of scores

$$d = \frac{M_1 - M_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{2}}}$$

$$r = \frac{d}{\sqrt{((D^2) + 4)}}$$

Power Calculators

G-Power

R Statsmodels

Python Statsmodels

Computing the Sample Size for T test

```python
# import required modules
from math import sqrt
from statsmodels.stats.power import TTestIndPower

#calculation of effect size
# size of samples in pilot study
n1, n2 = 4, 4

# variance of samples in pilot study
s1, s2 = 5**2, 5**2

# calculate the pooled standard deviation
# (Cohen's d)
s = sqrt(((n1 - 1) * s1 + (n2 - 1) * s2) / (n1 + n2 - 2))

# means of the samples
u1, u2 = 90, 85

# calculate the effect size
d = (u1 - u2) / s
print(f'Effect size: {d}')

# factors for power analysis
alpha = 0.05
power = 0.8

# perform power analysis to find sample size
# for given effect
obj = TTestIndPower()
n = obj.solve_power(effect_size=d, alpha=alpha, power=power,
                    ratio=1, alternative='two-sided')

print('Sample size/Number needed in each group: {:.3f}'.format(n))
```

```
Effect size: 1.0
Sample size/Number needed in each group: 16.715
```

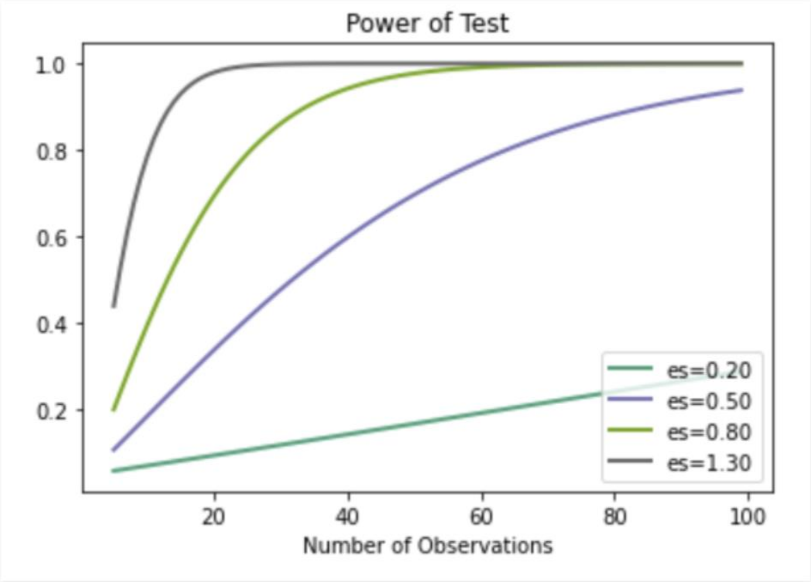
Computing the Power for T Test

```
from statsmodels.stats.power import TTestPower


power = TTestPower()
n_test = power.solve_power(nobs=40, effect_size = 0.5,
                           power = None, alpha = 0.05)
print('Power: {:.3f}'.format(n_test))
```

```
Power: 0.869
```

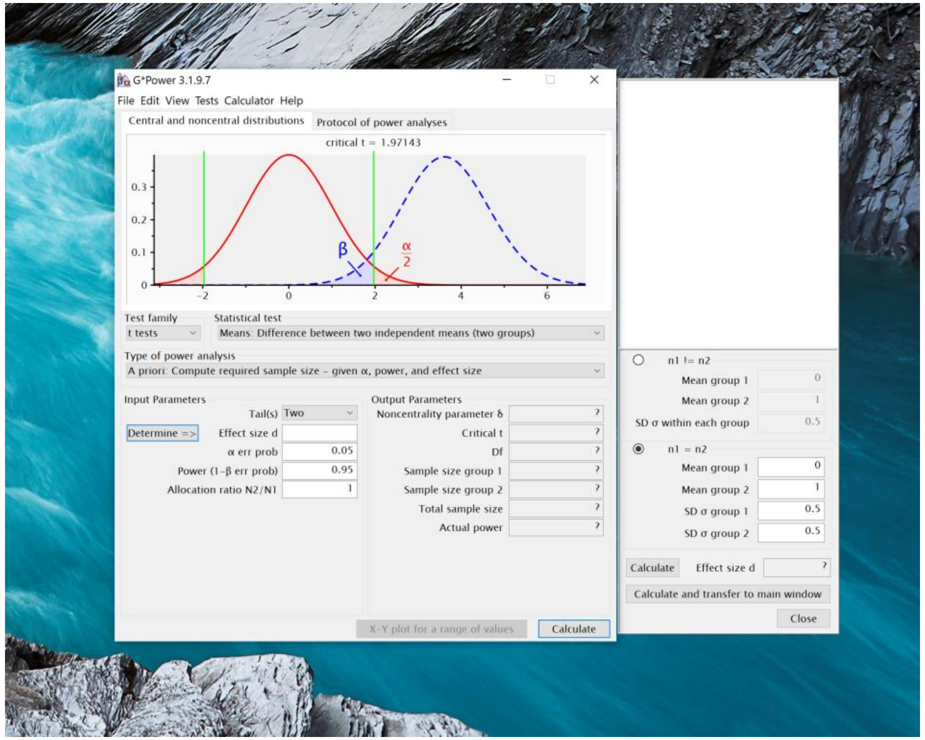
Samples vs. Power for different effect sizes

```python
# import required libraries
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.stats.power import TTestIndPower


# power analysis varying parameters
effect_sizes = np.array([0.2, 0.5, 0.8,1.3])
sample_sizes = np.array(range(5, 100))

# plot power curves
obj = TTestIndPower()
obj.plot_power(dep_var='nobs', nobs=sample_sizes,
               effect_size=effect_sizes)

plt.show()
```



G-Power

Universität Düsseldorf



Other resources

Preregistration and power analysis:

Best Practices for Transparent Social Science

Methods to tackle each potential problem

I. Uncommon

Cataloguing publications that need to be retracted

Disclosure of all data and materials

Fraud detection flags

II. Common, but not primary culprit

Pre-registration & Journal acceptance

Statistical checks

Power analysis

III. Very common

Meta-Statistics

(Pre-registration/Registered reports)

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

P-Curve: A Key to the File Drawer”

Simonsohn, Nelson and Simmons (2014)

Focus on the distribution of p-values < .05

Look at “evidential value” of “form of” p-hacking

Empirical simulation

P-Curve: A Key to the File Drawer

Tests are more likely to be published when they are statistically significant

P-Curve: A Key to the File Drawer

Tests are more likely to be published when they are statistically significant

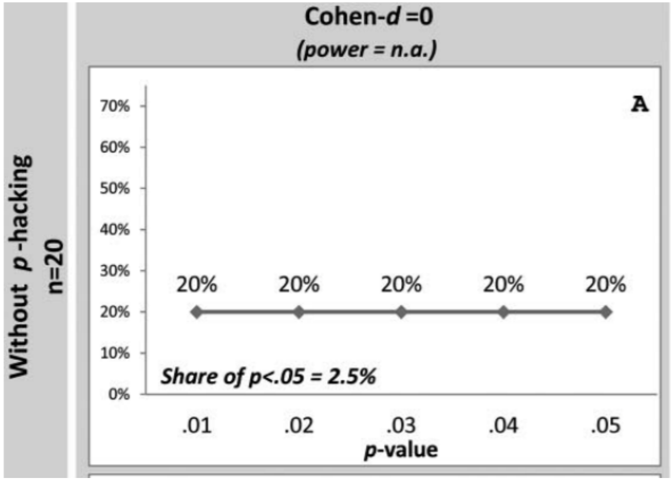
P-curve can test for presence or lack of evidential value but not prove that the theory is supported

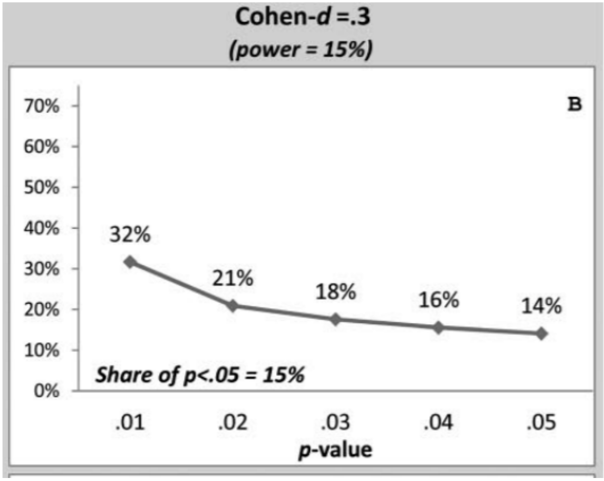
P-Curve: A Key to the File Drawer

Tests are more likely to be published when they are statistically significant

P-curve can test for presence or lack of evidential value but not prove that the theory is supported

Uses only p-values < .05

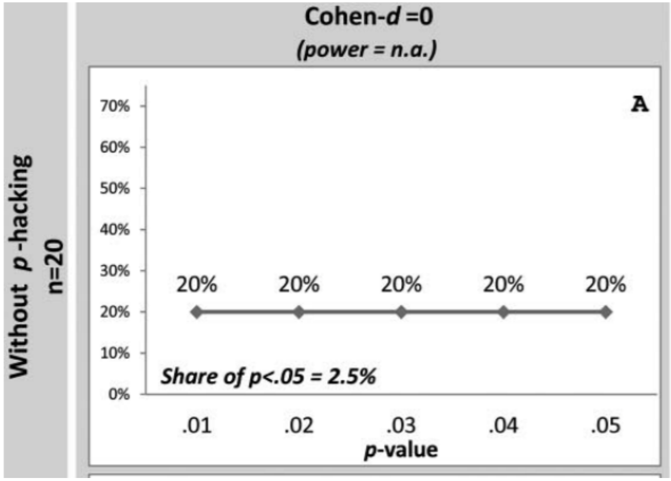


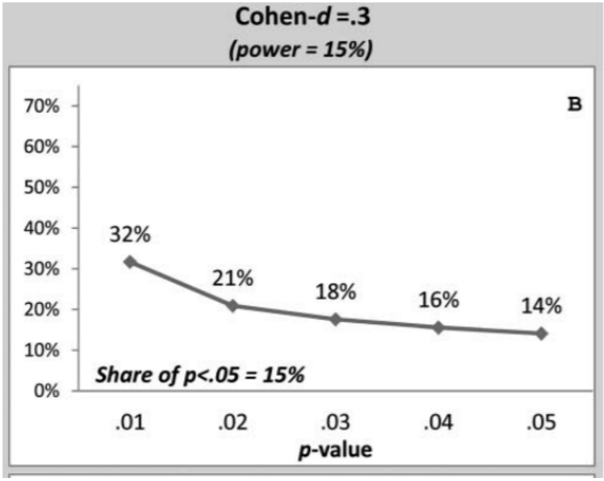


Distribution of P-values under “no effect” (d=0)

-> Uniform Distribution

Distribution of P-values with an effect (d>0) -> Right-skewed distribution



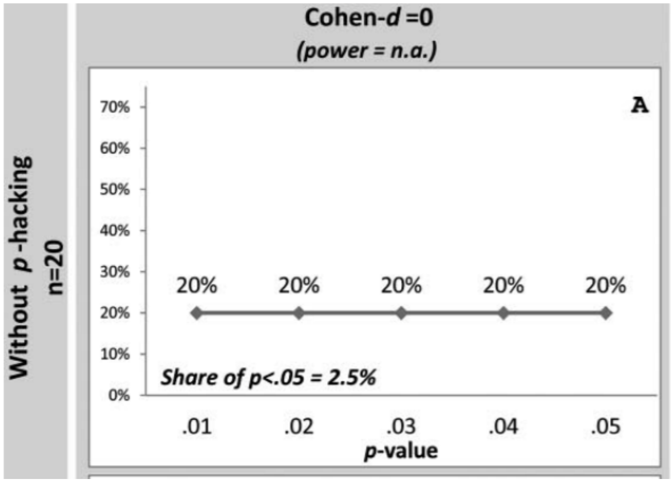


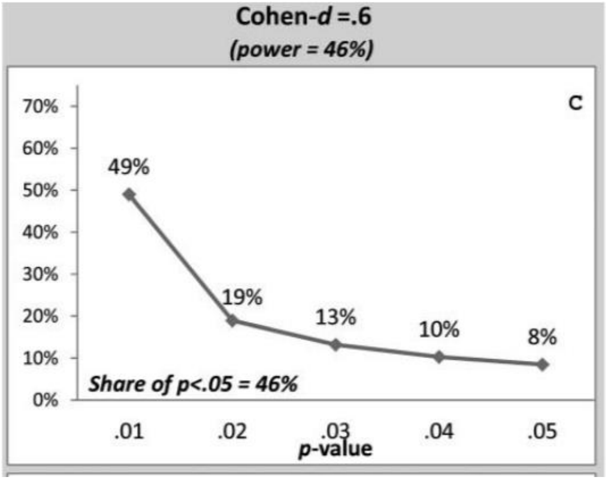
What happens when we increase power?

Distribution of P-values under “no effect” (d=0)

-> Uniform Distribution

Distribution of P-values with an effect (d>0) -> Right-skewed distribution

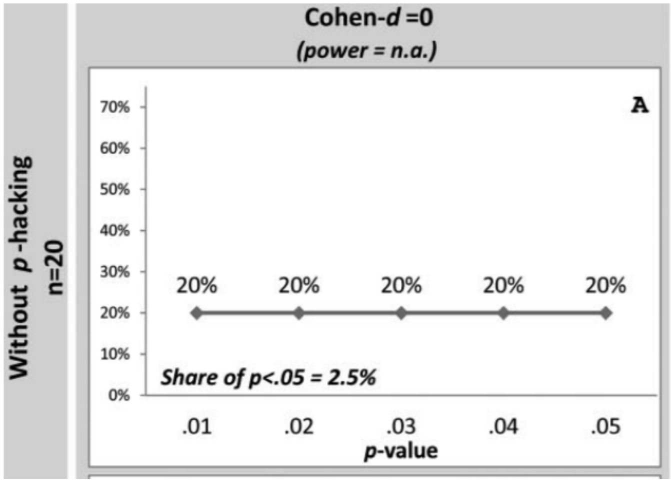


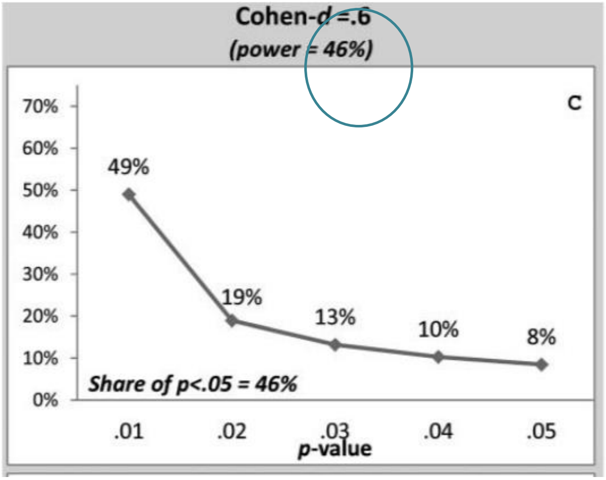


Distribution of P-values under “no effect” (d=0)

-> Uniform Distribution

Distribution of P-values with an effect (d>0) -> Right-skewed distribution

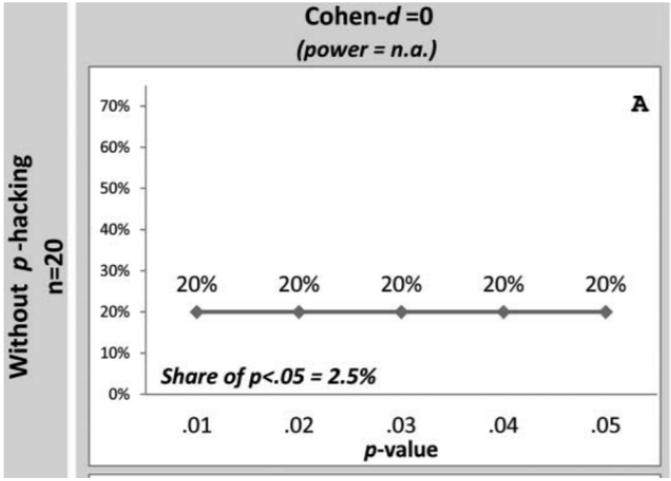


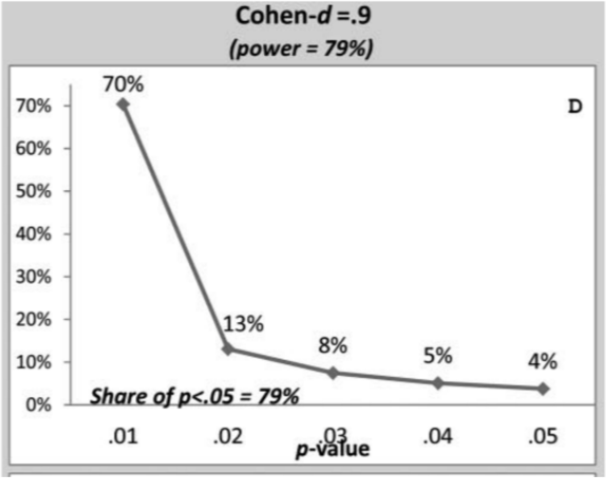


Distribution of P-values under “no effect” (d=0)

-> Uniform Distribution

Distribution of P-values with an effect (d>0) -> Right-skewed distribution





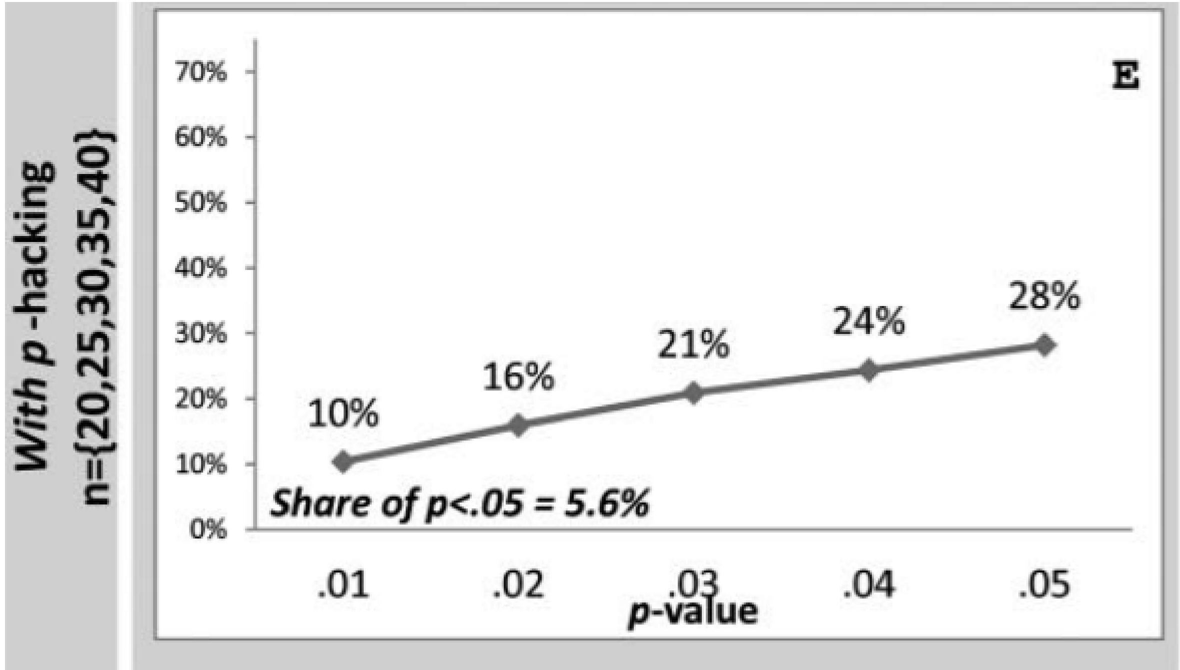
Distribution of P-values under "no effect" (d=0)

-> Uniform Distribution

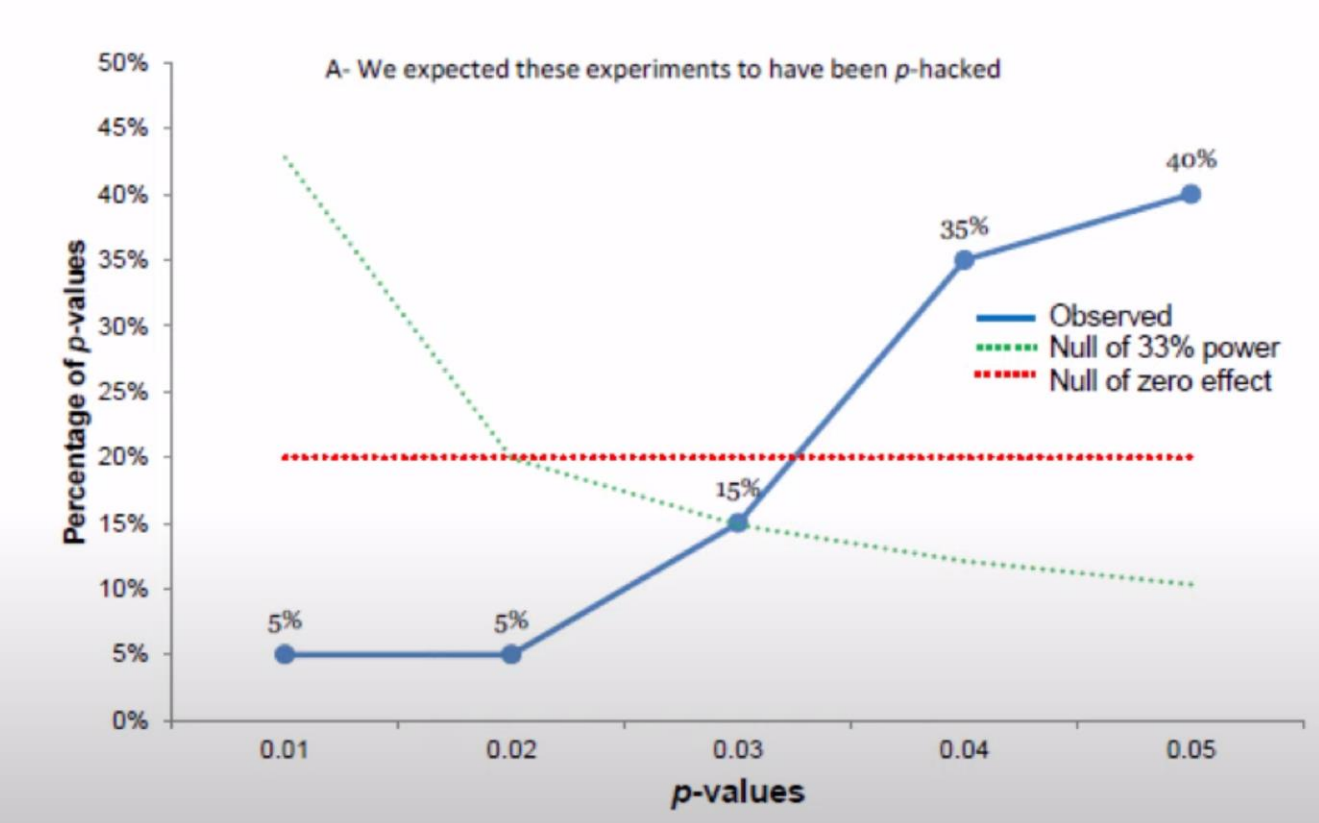
Distribution of P-values with an effect (d>0) -> Right-skewed distribution

What happens with p-hacking

P curve where they applied an early stopping rule (p-hacking)



P-curve of a psychology journal with suspected p-hacking



Answers questions

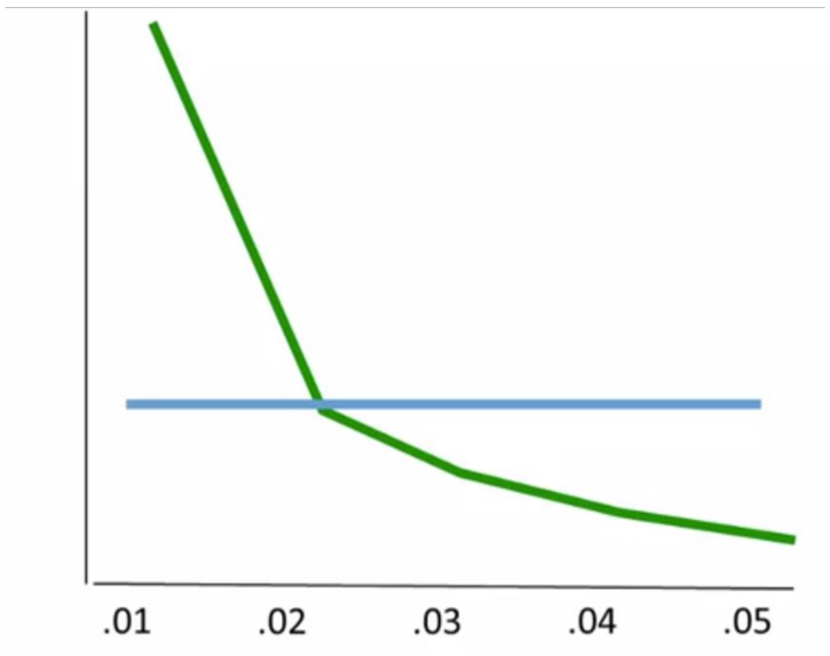
A) Does the p-value look like one where there is an effect or there is no effect? (**right-skew**)

- Compute termed ‘pp value’ with **null**
- Use Fisher’s method on pp values

B) Is there enough power to detect an effect from this literature?

- Compute ‘pp value’ with **33% power**
- Use Fisher’s method on pp values

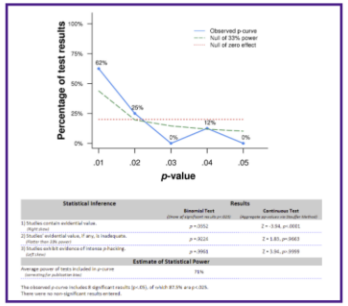
C) ‘Half curve’ formulation with $p < .025$

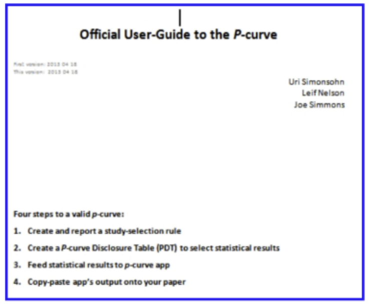


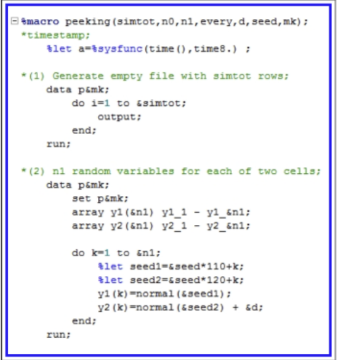
How to conduct a p-curve analysis for homework

P-curve.com

Paper A Evidential Value	Paper 2 Effect size	Paper 3 'Better P-curves' (robustness)	The online app 4.0	The User Guide	Supp Materials

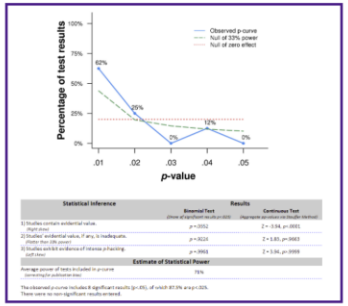
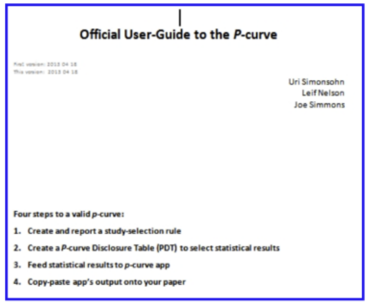
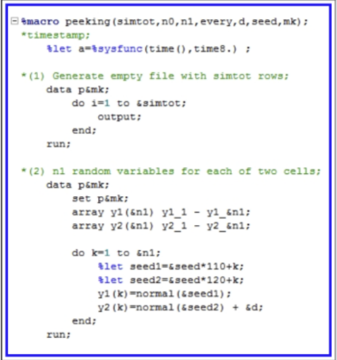






How to conduct a p-curve analysis for homework

P-curve.com

Paper A Evidential Value	Paper 2 Effect size	Paper 3 'Better P-curves' (robustness)	The online app 4.0	The User Guide	Supp Materials
					

P-curve app

p-curve app 4.06
How has the app changed? See summary.
Highlights of user guide
1) Not all p-values in a paper are selected, only those testing hypothesis of interest (See Table 3 in paper/user-guide).
2)In a 2x2 experimental design:
If an effect is predicted to attenuate, the p-value of the interaction is selected.
If an effect is predicted to reverse, the p-value of both simple effects are selected.
3) If you make a p-curve public, report a P-curve Disclosure Table (see Table 2 in paper/user-guide for an example).
Questions about p-curve? Email Uri, Leif, or Joe.
Enter your tests:
Go ahead. Replace the examples.
t(88)=2.1
r(147)=.246
F(1,100)=9.1
f(2,210)=4.45
Z=3.45
chi2(1)=9.1
r(77)=.47
chi2(2)=8.74
Make the p-curve

P-curve guidelines

Step 1. Create a study-selection rule

P-curve can be used to assess the evidential value of diverse sets of findings.

If a rule can be specified that creates a meaningful set of studies, then *p*-curve can validly assess the set's joint evidential value.

The rule should be set in advanced, before statistical results are analyzed, and disclosed in the paper.

Examples of rules:

- The yearly top-5 most cited articles in the *Quarterly Research Journal* 1984-1989
- All studies published in 2009 with wine as a manipulation and simulated driving behavior as a dependent variable.
- The most recent 10 articles published by proctologist Giordano Armani.
- Clinicaltrials.gov registered studies examining antidepressants among teenagers.

https://www.p-curve.com/guide.pdf

P-curve guidelines

Step 2. Create a *P*-curve Disclosure Table to select results to analyze

Table 1 summarizes the steps for creating a disclosure table.
Table 2 provides an example.

Table 1.Five Steps to Create a *P*-curve Disclosure Table

Step 1	Identify researchers' stated hypothesis and study design quoting from paper	(Columns 1 &2)
Step 2	Identify the statistical result testing stated hypothesis using Table 3	(Column 3)
Step 3	Report the statistical results of interest quoting from paper	(Column 4)
Step 4	Recompute the precise p -value(s) based on reported test statistics	(Column 5)
Step 5	Report robustness results	(Column 6)

Table 3 in paper. Which statistical result to select for *p*-curve?

DESIGN	EXAMPLE	WHICH RESULT TO INCLUDE IN MAIN P-CURVE	WHICH RESULT TO INCLUDE IN ROBUSTNESS TEST
3-Cell	*Examining how math training affects math performance*		
High **Medium** **Low**	60 minutes of math training 30 minutes of math training 5 minutes of math training	Linear trend	High vs. low comparison
Treatment **Control 1** **Control 2**	60 minutes of math training 60 minutes of unrelated training No training	Treatment vs. Control 1	Treatment vs. control 2
Treatment 1 **Treatment 2** **Control**	60 minutes of math training, start with easy questions 60 minutes of math training, start with hard questions No training	Treatment 1 vs. Control	Treatment 2 vs. Control
2X2 DESIGN	*Examining how season interacts with being indoors vs. outdoors to affect sweating*		
Attenuated Interacton	Always sweat more in summer, but less so when indoors.	2x2 Interaction	
Reversing Interacton	Sweat more in summer than winter when outdoors, but more in winter than in summer when indoors	Both simple effects	
3x2 DESIGN	*Examining how season interacts with math training to affect math performance*		
Attenuated Trends	More math training (60 vs. 30 vs. 5 minutes) leads to better performance always, but more so in winter than in summer	Difference in linear trends	2x2 interaction for high vs. low
Reversing Trends	More math training (60 vs. 30 vs. 5 minutes) leads to better performance in winter, but worse performance in summer	Both linear trends	Both high vs. low comparisons
2x2x2 DESIGN	*Examining how gender and season interact with being indoors vs. outdoors to affect sweating*		
***Attenuation* of attenuated interaction**	Both men and women sweat more in summer than winter, but less so when indoors. This attenuation is stronger for men than for women.	Three-way interaction	
***Reversal* of attenuated interaction**	Men sweat more in summer than winter, but less so when indoors. Women also sweat more in summer than winter, but more so when indoors.	Both two-way interactions	
***Reversal* of reversing interaction**	Men sweat more in summer than winter when outdoors, but more in winter than in summer when indoors. Women sweat more in winter than summer when outdoors, but more in summer than winter when indoors.	All four simple effects	

Keep in mind:

Important!

1. **In a 2x2 design,**
   - If attenuation is predicted, select only the interaction
   - If a reversal is predicted, select only both simple effects

2. ***Discrete tests*.**
   P-curve is only approximately valid for discrete tests (e.g., comparing proportions). *P*-curves of discrete tests are, for now, merely suggestive. See Supplement #4.

P-curve guidelines

To check heterogeneity in your estimate, use R package dmetar.pcurve*

https://dmetar.protectlab.org/reference/pcurve.html

Step 3. Feed key results to *p*-curve app (version 3.0)

The web-based app looks like this:

You can copy paste your tests in the format used in the examples there.
If you have results p>.05, the app will automatically exclude them and report how many were excluded.

P-curve guidelines

Step 4: Report all output on paper

Problems with P-curve

Heterogeneity of effect sizes

Can’t use with tests of discrete data (using Chi Square test, F test)

Interpreting the average power and effect size of the estimate is problematic

- Average Power: A Cautionary Note (McShane et al.)

Disclosure of studies is very important

- Negative Effect of a Contractive Pose Is Not Evidence for the Positive Effect of an Expansive Pose: Commentary on Cuddy, Schultz, and Fosse (2018)

Categorical sin of P values (professor priming research)

- Professor Priming discussion

Problems with P-Curve

Gelman take from blog:

"McShane et al. and Simonsohn et al. that these methods should be thought of as methods of demonstrating **how bad the selection bias can be in a literature, under best-case assumptions**, rather than as a method of estimating underlying effect sizes.

Thus, **I can see how the observed distribution of p-values can be helpful to look at**, if for no other reason than to **reveal problems with naive interpretations of published p-values**"

Problems with P-curve

Gelman take from blog:

" general view that all these tools are most useful as a sort of rhetorical approach to show how bad things can be, even in the best-case scenario.

I get concerned, though, if people take these methods too literally. Consider the classic file-drawer-effect paper by Rosenthal, which I assume was written to demonstrate how serious this selection problem can be, but is sometimes twisted around to give the opposite meaning (by doing the calculation of how many papers would need to have been discarded to be consistent with a particular pattern of published results, and then claiming that since no such massive "file drawer" exists, the published claims should be accepted). I wouldn't want researchers to take p-curve, or the Hedges approach, as evidence that a literature of uncontrolled p-values is approximately just fine.

As is often the case, **I find myself more convinced by the demonstration of bias than by the attempted bias correction**. In that sense, I see the Hedges procedure, or p-curve, or p-uniform, as being comparable to Type M and Type S errors (Gelman and Tuerlinckx, 2000) as a way of quantifying some effects of selection bias in statistical inference, but the desired solution is to go back to the original, unselected, data. All these methods can be useful in giving us a sense of the scale of bias arising in idealized situations.

"

Other meta-analytic estimates to supplement when seeing right-skew

Z-curve
  https://zcurve.shinyapps.io/zcurve19/

Selection procedure (Hedges-G)

Funnel Plot (Trim and Fill Method)

For a comprehensive review of publication bias, **highly recommend**:
  Doing Meta Analysis in R (Harrer et al.)
  https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/pub-bias.html

Summary

Factors leading to replication failures

I. Uncommon

Fraud

II. Common, but not primary culprit

File-drawering failed studies

Innocent Errors

Insufficient power

III. Very common

P-hacking

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Methods to tackle each potential problem

I. Uncommon

Cataloguing publications that need to be retracted

Disclosure of all data and materials

Fraud detection flags

II. Common, but not primary culprit

Pre-registration & Journal acceptance

Statistical checks

Power analysis

III. Very common

Meta-Statistics

(Pre-registration/Registered reports)

False-Positive Psychology – Undisclosed flexibility in data collection and analysis allows presenting anything as significant Simmons et al. 2011

Summary

Multiple methods have emerged to deal with these problems but they still have **limitations**

Registered reports and their increased acceptance along with **well powered research designs** based on curated findings (replicated) may be good path forward now

Homework assignment discussion

Questions?