

Anne Gärtner
Faculty of Psychology

Workshop Open Science Practices

Part 1

P-hacking

Overview

Time

13:00 – 13:10
13:10 – 13:40
13:40 – 13:55
13:55 – 14:40
14:40 – 14:55
14:55 – 15:40
15:40 – 16:10
16:10 – 16:25
16:25 – 16:40
16:40 – 17:25
17:25 – 17:40
17:40 – 18:00

Topic

Welcome
Power (Alex)
Discussion
Data Collection (Anne)
Break
P-hacking (Anne)
Publication Bias (Alex)
Discussion
Break
Preregistration (Anne)
Discussion
Wrap Up, Evaluation

Workshop material

- ▲  MGK Open Science Module
- ▲  Registration
-  Introduction
- ▶  W1 - Good Scientific Practice
- ▶  W2 - Research Data Management
- ◀  W3 - Research Transparency
-  General Information
-  0. Introduction
-  1. Open Science
-  2. Open Access
-  3. Open Data, Materials, and Co
-  4. Reproducible Analyses
-  5. Preregistration
-  Opt.: Replication Research
-  Workshop Slides
-  Literaturverzeichnis

Zoom Poll



Outline

What is *p*-hacking?

How to *p*-hack: Tools

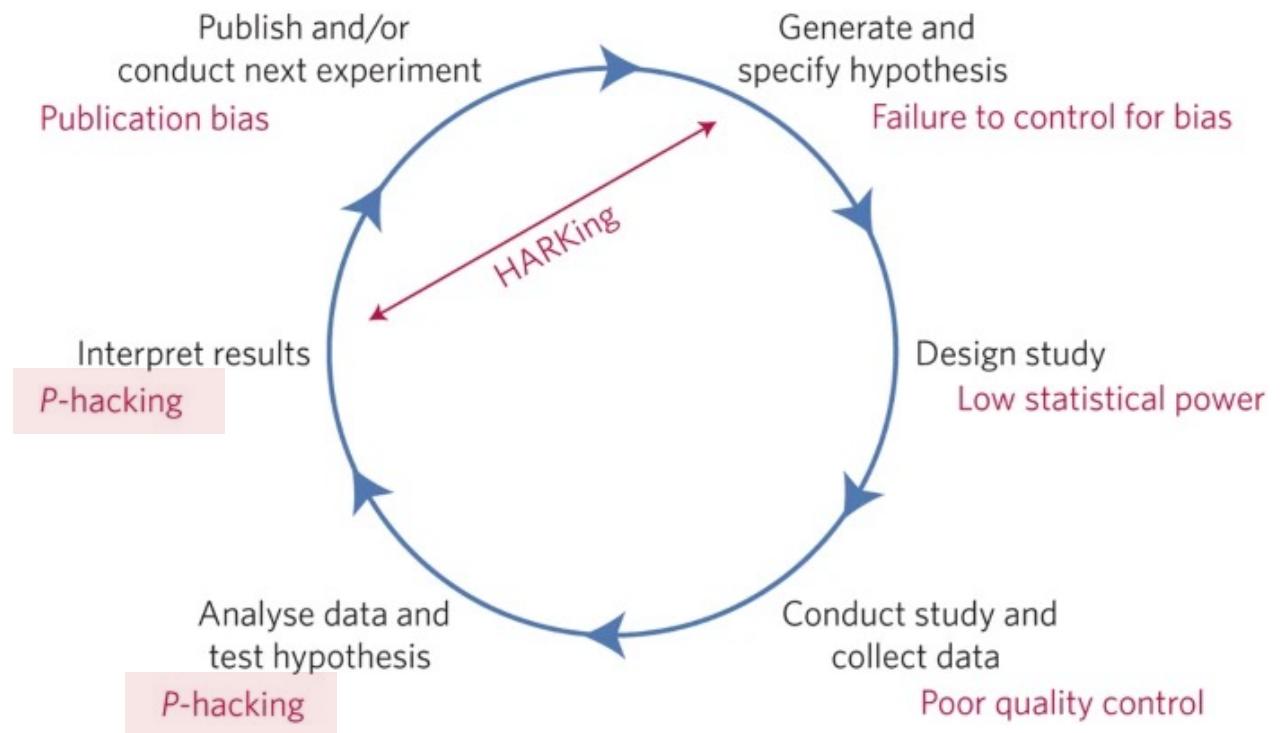
- Outcome switching, selective reporting, optional stopping, subgroup analysis, flexible measures, no preregistration

Exercise

- Train with *p*-hacker

Discussion

Solutions and further recommendations



What is *p*-hacking?

Introduction

A tale of two papers (by Michael Inzlicht)

Original version

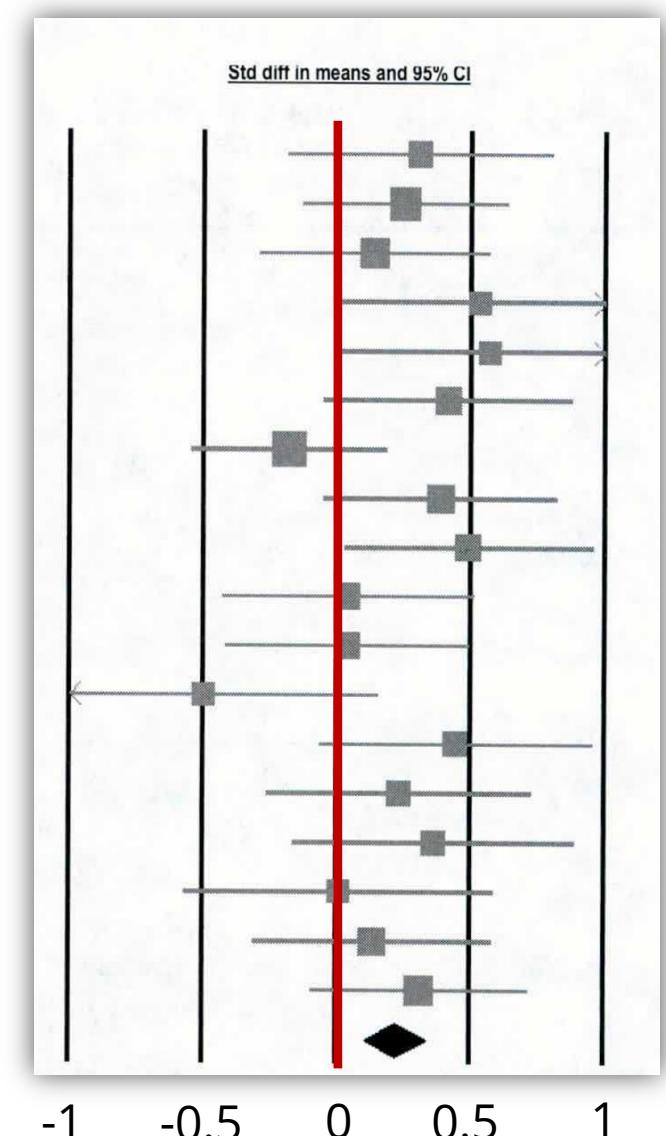
- 7 experiments
- 7/7 significant
- effect sizes in the medium to large range
- ad-hoc covariates
- “Excessive significance”

“The first [paper] was emblematic of the old way of doing business, with 7 studies that were scrubbed clean to be near-perfect.”

Revised version

- 18 experiments
- 2/18 significant
- Some studies with reversed direction

“This is what real data look like. The data are not always pretty, they have warts, but they are real.”



Introduction

How to become a Professor

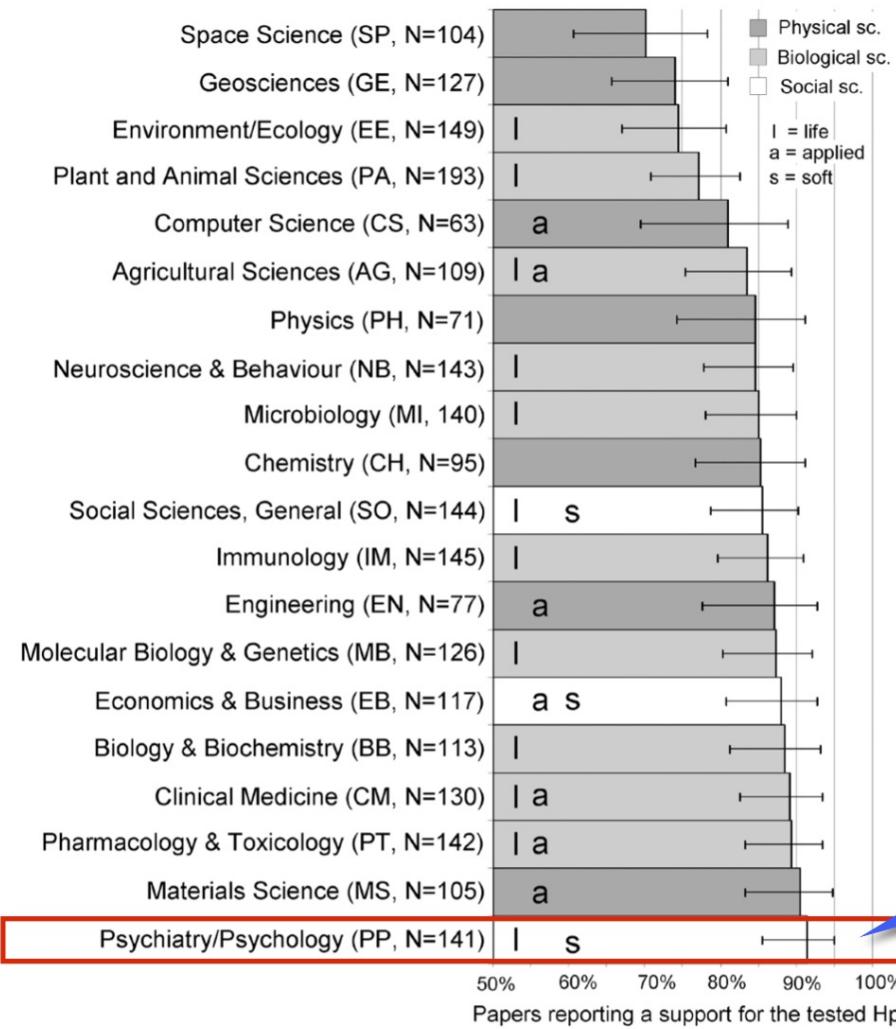
Actual (not desired) relevance in professorship hiring committees	Rank
Number of peer-reviewed publications	1
Fit of research profile to the hiring department	2
Quality of research talk	3
Number of publications	4
Volume of acquired third-party funding	5
Number of first authorships	6
...	...

N = 1453 psychology researchers, 66% were actually members of a professorship hiring committee

Introduction

How to get lots of publications

“Positive” results increase down the hierarchy of the sciences



92% of published papers have significant, positive results

What is *p*-hacking?

Researchers are not rewarded for being right,
but rather for publishing a lot.

Nelson, Simmons, & Simonsohn (2012); Nosek, Spies, Motyl (2012); Munafò (2016)

Shit Academics Say
@AcademicsSay

A. I get paid to think.
B. About what.
A. Tenure mostly.

[Tweet übersetzen](#)

5:05 vorm. · 14. Jan. 2015 · Twitter for iPhone

172 Retweets **271 „Gefällt mir“-Angaben**

What is *p*-hacking?

***p*-hacking (n.).** Tune your data analysis in a way that you achieve a significant *p*-value in situations where it would have been non-significant.

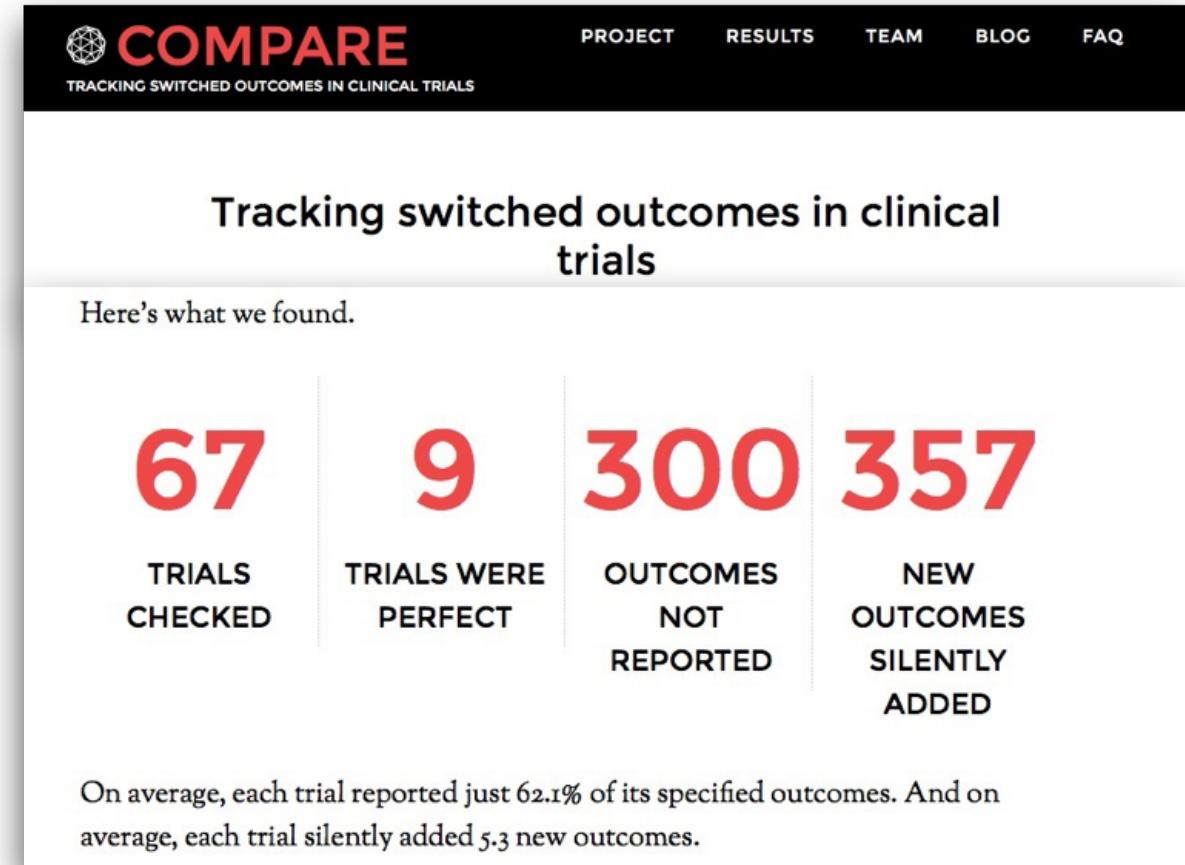
Questionable research practices (QRPs) (n.).
Practices of data collection and data analysis that are not outright fraud, but also not really kosher.

p-hack your way to scientific glory!

How to *p*-hack: Tools

Tool 1: Outcome switching

- Assess more than one dependent variable (DV), but report only those which “worked”
- 2 outcome variables: false positive rate increases from **5%** → **9.5%**
- 5 (uncorrelated) DVs with one-sided testing: false positive rate increases from **5%** → **41%**
- How prevalent is it?
 - **66% of researchers** admit having done this (John, Loewenstein, & Prelec; 2012)



How to *p*-hack: Tools

Tool 2: Many conditions

- Assess more than two conditions but report only those with $p < .05$
- *Example 1:* testing “high”, “medium”, and “low” conditions and reporting only the results of a “high” versus “medium” comparison
- *Example 2:* Thesis vs. manuscript (O’Boyle et al., 2017)
- Gives you more than one chance to find an effect, increases false positive rate up to → **12.6%**
- How prevalent is it?
 - **27% of researchers** admit having done this (John et al., 2012)



Joe Hilgard
@JoeHilgard

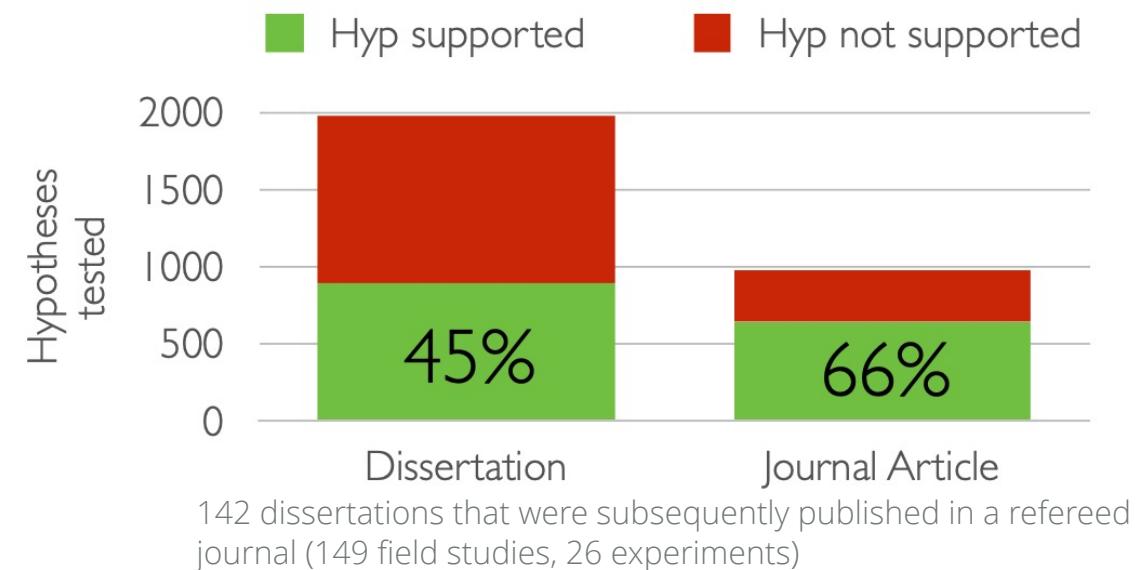
Here's another spicy one: Thesis reports four conditions, 415 subjects. Manuscript reports three conditions, 140 subjects.



Joe Hilgard @JoeHilgard · 16. Feb.

Figured it out: It started with a $2 \times 2 \times 4$ design and worked its way down to the 2×3 design that “worked.”

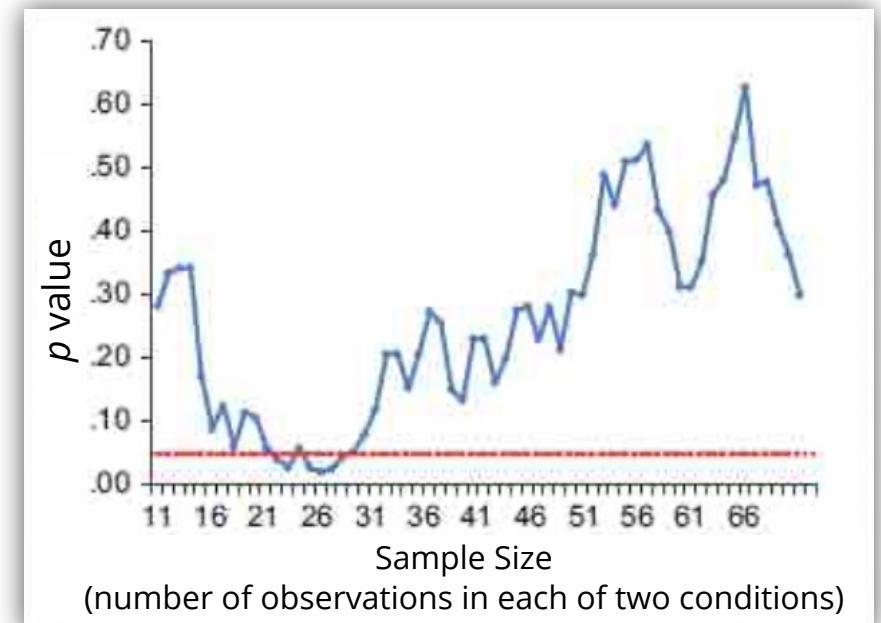
The Chrysalis effect: How ugly initial results metamorphosize into beautiful articles



How to *p*-hack: Tools

Tool 3: Optional stopping

- Collect an initial sample, analyze results, add additional participants if not significant, repeat until significance is found
- Increase once: false positive rate = **7.7%**, twice: false positive rate = **11%**
- But with enough looks can be pushed to → **100%!**
- How prevalent is it?
 - **70% of researchers** admit having continued or stopped data collection based on looking at the interim results (John et al., 2012)



How to *p*-hack: Tools

Tool 4: Subgroup analysis

- Research question: Do aggressive primers trigger aggressive behavior?

A second study in Turner, Layton, and Simons (1975) collects a larger sample of men and women driving vehicles of all years. **The design was a 2 (Rifle: present, absent) × 2 (Bumper Sticker: "Vengeance", absent) design with 200 subjects.**



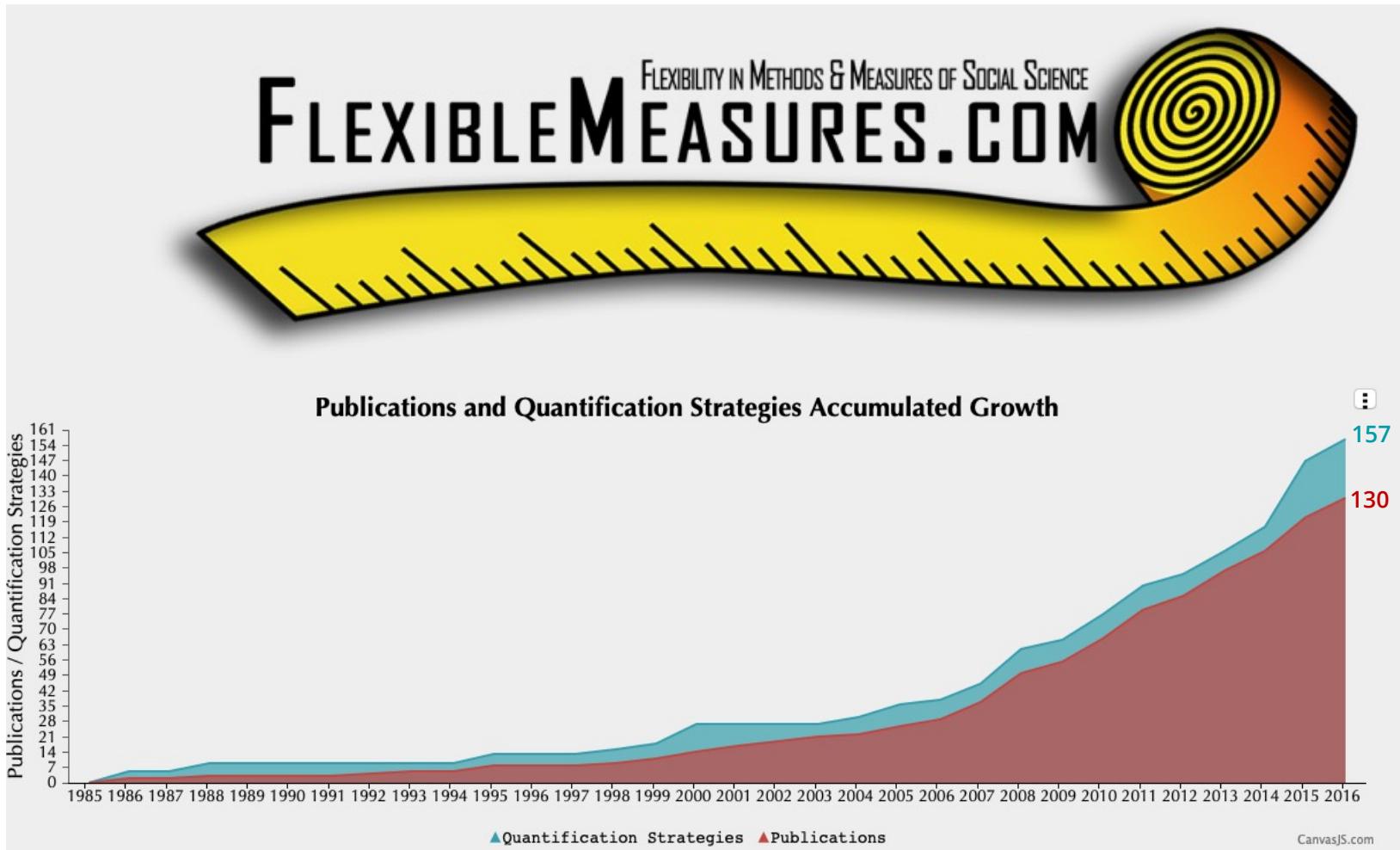
- Presumably no effect, but wait...

They **divide this further by driver's sex** and by a **median split on vehicle year**. They find that the Rifle/Vengeance condition increased honking relative to the other three, but only among newer-vehicle male drivers, $F(1, 129) = 4.03, p = .047$. But then they report that the Rifle/Vengeance condition decreased honking among older-vehicle male drivers, $F(1, 129) = 5.23, p = .024$! No results were found among female drivers.



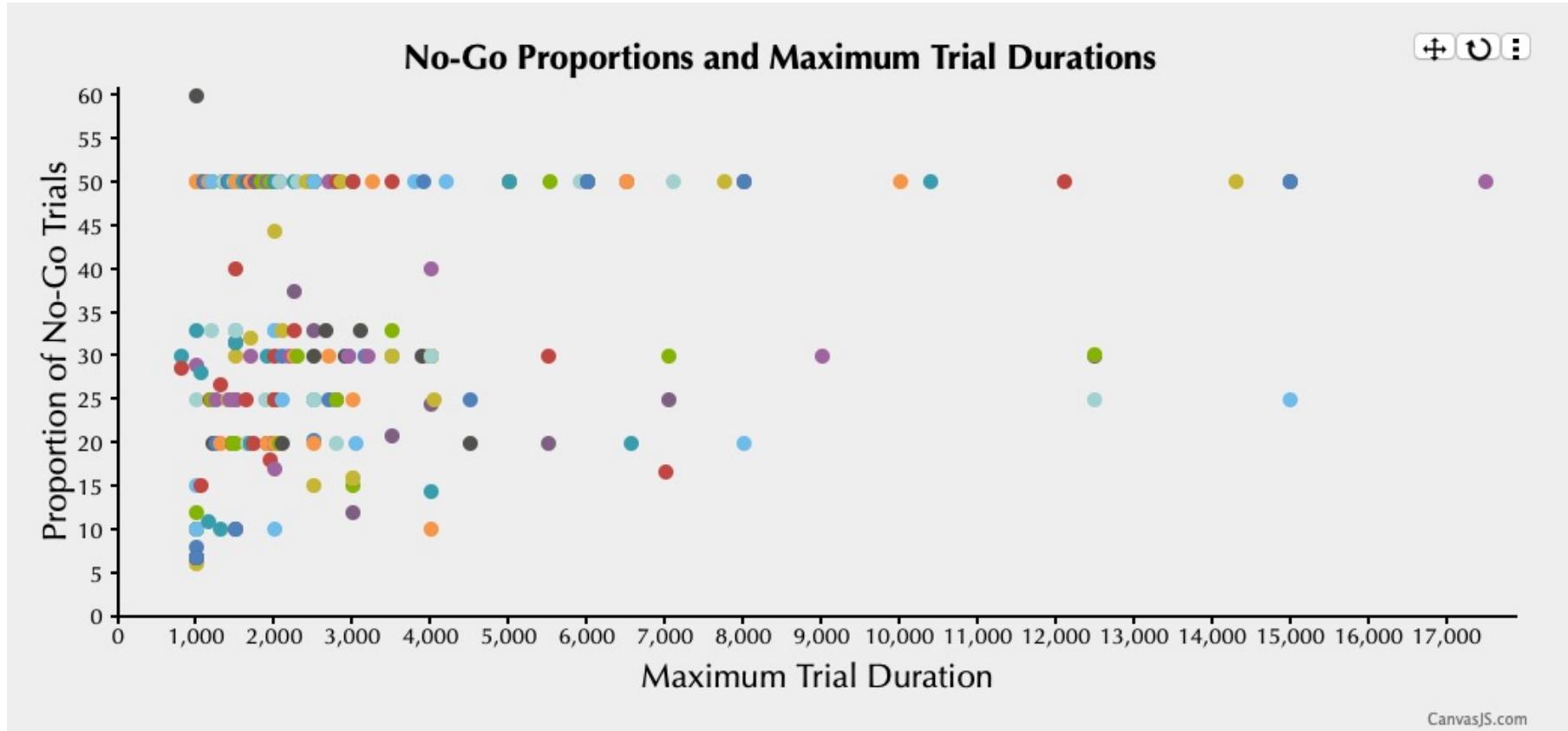
How to *p*-hack: Tools

Tool 5: Flexible measures



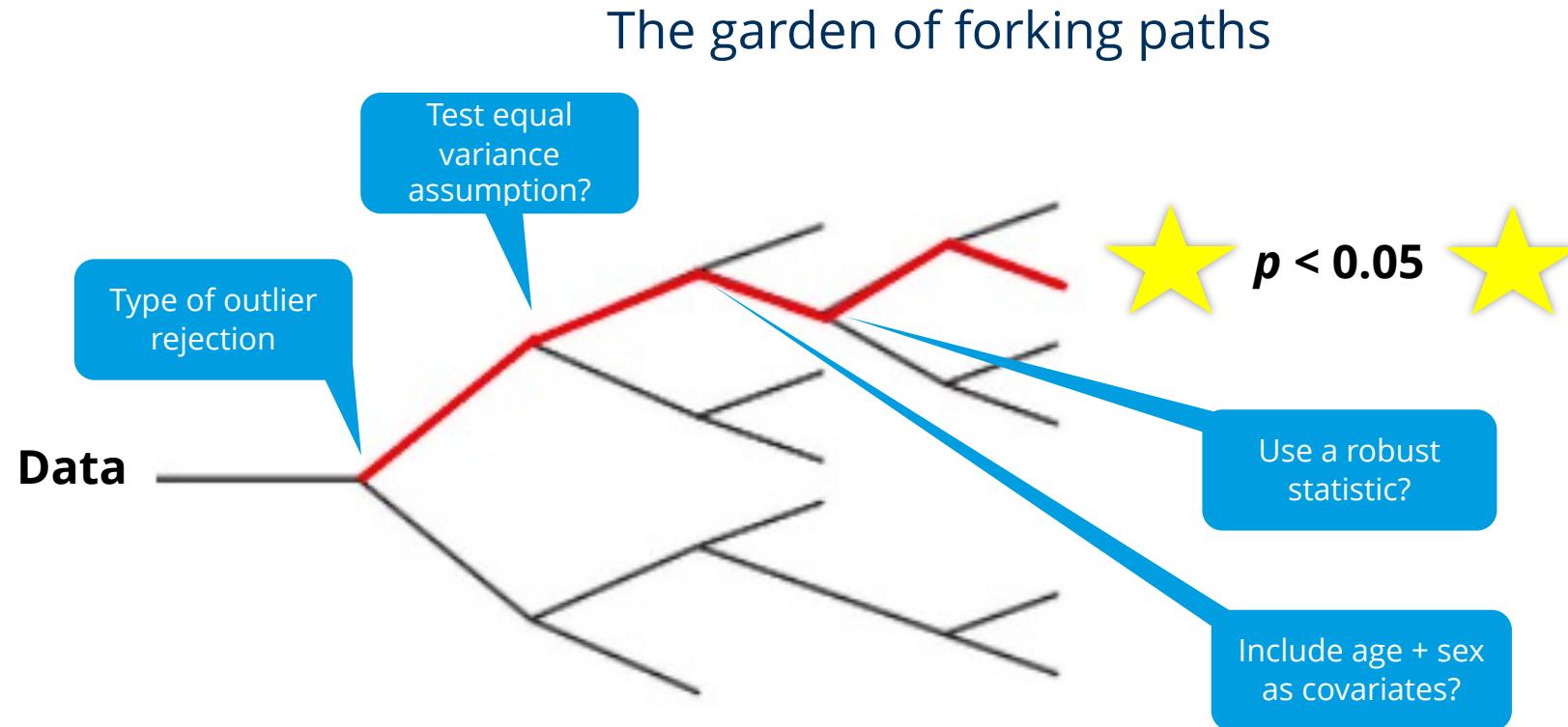
How to *p*-hack: Tools

Tool 5: Flexible measures



How to *p*-hack: Tools

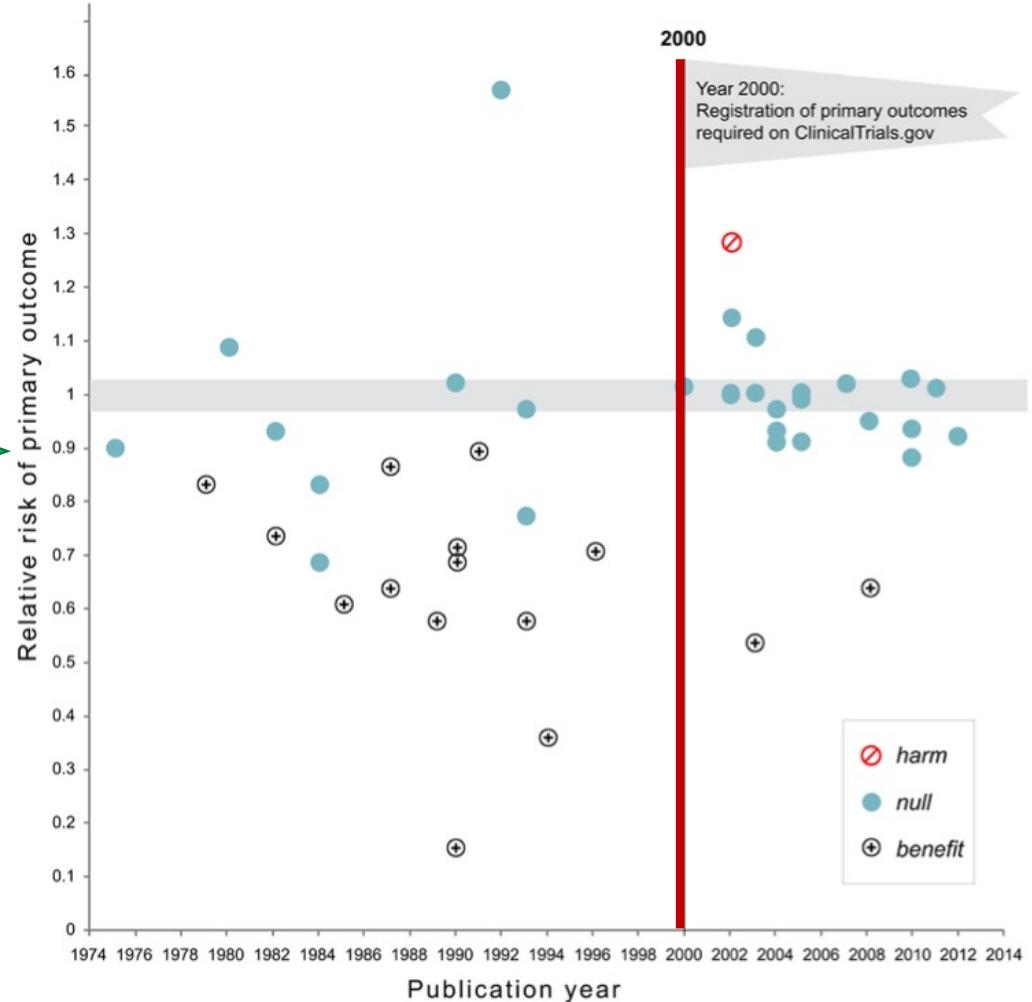
Tool 6: Many researcher degrees of freedom



How to *p*-hack: Tools

Tool 7: No preregistration

No prereg:
57%
success rate!



Prereg:
8%
success rate...

Exercise

Exercise

p-hacker: Train your p-hacking skills!

Manual Technical Details

New study Now: p-hack!

No study run yet - click on 'Run new experiment' at the bottom of the left panel!

Settings for initial data collection:

Name for experimental group
PhD SFB

Name for control group
PhD noSFB

Initial # of participants in each group
20

True effect (Cohen's d)
0

Number of DVs
4

Run new experiment

Go to <https://shinyapps.org/apps/p-hacker/> and try to p-hack the data!

<https://shinyapps.org/apps/p-hacker/>

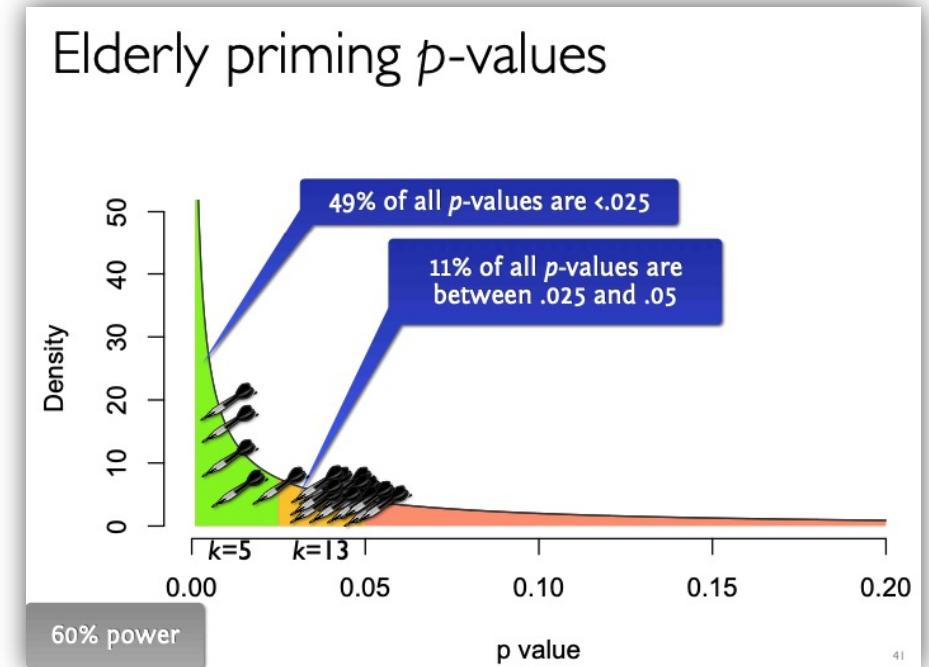
P-hacking

What we did *not* consider

Tools to detect *p*-hacking

- *P*-curve
- Replication Index (R-Index)
- Test of insufficient variance (TIVA)
- Begg and Mazumdar test
- Egger's regression test
- Precision effect test (PET)
- Precision effect estimate with standard error (PEESE)
- ...

→ Practice with *P*-checker!



p-checker The one-for-all *p*-value analyzer

Enter test statistics here:

```
# Easy mode: only enter the test statistics with c  
t(47) = 2.1  
chi2(1) = 9.15  
r(77) = .47  
F(1, 88) = 9.21  
p = .02  
p(48) = .018  
  
# add reported p-value; mark one-tailed; set alpha  
t(123) = 2.54; p < .01  
Z = 1.9; one-tailed; p=.03  
r(25) = 0.21; crit=.10
```

enü anzeigen paper_ID

Demo data by Slartibartfast
Go and replace the examples in the text box! # starts a comment
<http://shinyapps.org/apps/p-checker/>

Excess Significance TIVA p-Curve Meta-analysis p values correct?

R-Index analysis:

Success rate = 0.9286
Median observed power = 0.6575
Inflation rate = 0.2711
R-Index = 0.3864

For information about R-index, see <http://www.r-index.org/>.

Detailed results for each test statistic

Solutions

Solutions

How to prevent *p*-hacking

Be honest / Don't do intentional *p*-hacking

- 21 word solution (Simmons et al., 2012)

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

- Standard Reviewer Statement for the Disclosure of Sample, Conditions, Measures, and Exclusions

"I request that the authors add a statement to the paper confirming whether, for all experiments, they have reported all measures, conditions, data exclusions, and how they determined their sample sizes. The authors should, of course, add any additional text to ensure the statement is accurate. This is the standard reviewer disclosure request endorsed by the Center for Open Science [see <http://osf.io/hadz3>]. I include it in every review."

Solutions

How to prevent *p*-hacking

Don't do unintentional *p*-hacking

- Provide open material for an existing project
 - E.g. on OSF (<https://osf.io>), publish open data, open material, reproducible analysis code, preprints, postprints, supplemental material
 - Get a persistent URL and even a doi
 - Identify questionable research practices
 - E.g., *p*-checker, replication index (<https://replicationindex.com>)
- Preregister your studies

Reproducibility Project: Psychology

Contributors: Christopher Jon Anderson, Joanna Anderson, Marcel A.L.M. van Assen, Peter Raymond Attridge, Angela Attwood, Jordan Axt, Molly Babel, Stéphan Bahnik, Jennifer Beer, Raoul Bell, Heather Bentley, Don van den Bergh, Leah Beyan, Bobby den Bezemer, Denny Borsboom, Annick Bosch, Frank Bosco, Sara Bowman, Mark B Kristina Brown, Jovita Brunning, Ann Calhoun-Sauls, Shannon Callahan, Elizabeth Chagnon, Jesse J. Chandler, Christopher R. Chartier, Felix Cheung, Phuonguyen Chu, Li

Affiliated institutions: Laura and John Arnold Foundation, University of Virginia, Center For Open Science

Date created: 2012-04-01 05:49 PM | Last Updated: 2019-12-04 10:48 PM

Identifier: DOI 10.17605/OSF.IO/EZCUJ

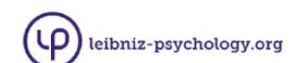
Category: Project

Description: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available.

License: CCO 1.0 Universal

Included in Metascience's Collection

The screenshot shows a detailed view of an OSF project. At the top, there's a header with the project name 'Reproducibility Project: Psychology' and its DOI. Below the header, there's a brief abstract about the project's goal of estimating reproducibility through replications. The main content area is titled 'Wiki' and contains the project's title and a link to 'Read More'. Below the wiki, there's a 'Files' section showing a list of files uploaded, including 'OSC2012.pdf' and several sub-folders like 'Analysis' and 'Replicator Resources'. To the right of the main content area, there's a sidebar titled 'Components' which lists various parts of the project such as 'Estimating the Reproducibility' (with a link to 'Nosek, Cohoon, Kidwell & 1 more'), 'Analysis' (with a link to 'Bakker, Borsboom, Bosco & 26 more'), 'Replicator Resources' (with a link to 'Nosek, Cohoon & Kidwell'), 'Presentations' (with a link to 'Nosek, Lai, LeBel & 4 more'), 'Post-Publication Additions and Corrections' (with a link to 'Cohoon & Kidwell'), 'Comments' (with a link to 'Kidwell'), and 'Replication of Janiszewski & Uy'.



Solutions

How to prevent *p*-hacking

Do post-publication peer review

Conceptualization of task boundaries preserves implicit sequence learning under dual-task conditions
Psychonomic Bulletin & Review (2013) - 1 Comment
pubmed: 23444106 doi: 10.3758/s13423-013-0409-0 issn: 1531-5320 issn: 1069-9384

Kimberly M. Halvorsen, Tana Truelove Wagschal, Eliot Hazeltine

#1 Statcheck commented September 2016

Using the R package statcheck (v1.0.1), the HTML version of this article was scanned on 2016-08-05 for statistical results (*t*, *r*, *F*, *Chi²*, and *Z* values) reported in APA format (for specifics, see Nuijten et al., 2015). An automatically generated report follows.

The scan detected 7 statistical results in APA format, of which 0 contained potentially incorrect statistical results, of which 0 may change statistical significance ($\alpha = .05$). Potential one-tailed results were taken into account when 'one-sided', 'one-tailed', or 'directional' occurred in the text.

Note that these are not definitive results and require manual inspection to definitively assess whether results are erroneous.

PubPeer browser plugin: Automatic alert for PubPeer comments at Google, journal websites etc.

- [Up-regulation of microRNA-10b is associated with the development of breast cancer brain metastasis.](#)
3. Ahmad A, Sethi S, Chen W, Ali-Fehmi R, Mittal S, **Sarkar FH**. *Am J Transl Res.* 2014 Jul 18;6(4):384-90. eCollection 2014.
PMID: 25075255 [Free PMC Article](#)



2 comments on PubPeer (by: Unregistered Submission)

[Similar articles](#)

Who to ask?

- Your local open science initiative
<https://tu-dresden.de/mn/psychologie/die-fakultaet/open-science>

OPEN SCIENCE INITIATIVE

OSIP Open Science Initiative
der Fakultät Psychologie der TUD

Interview with a researcher



<https://www.youtube.com/watch?v=ZaNtz76dNSI&sns=em>

Exercise

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY Republicans **Democrats**

2 DEFINE TERMS 3 IS THERE A RELATIONSHIP? 4 IS YOUR RESULT SIGNIFICANT?

Which politicians do you want to include?

Presidents
 Governors
 Senators
 Representatives

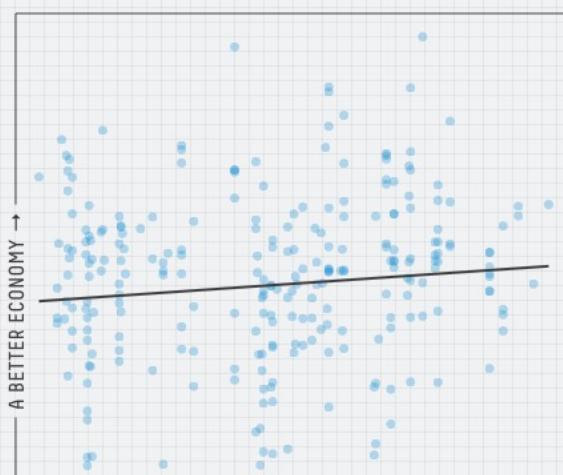
How do you want to measure economic performance?

Employment
 Inflation
 GDP
 Stock prices

Other options

Factor in power

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.



If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Almost

Your **0.10 p-value** is close to the **0.05 threshold**. Try tweaking your variables to see if you can push it over the line!

<https://projects.fivethirtyeight.com/p-hacking/>

Thank you!