Learn to *p*-hack like the pros!

J.J. at the English language Wikipedia

Dr. Felix Schönbrodt
Ludwig-Maximilians-Universität
München

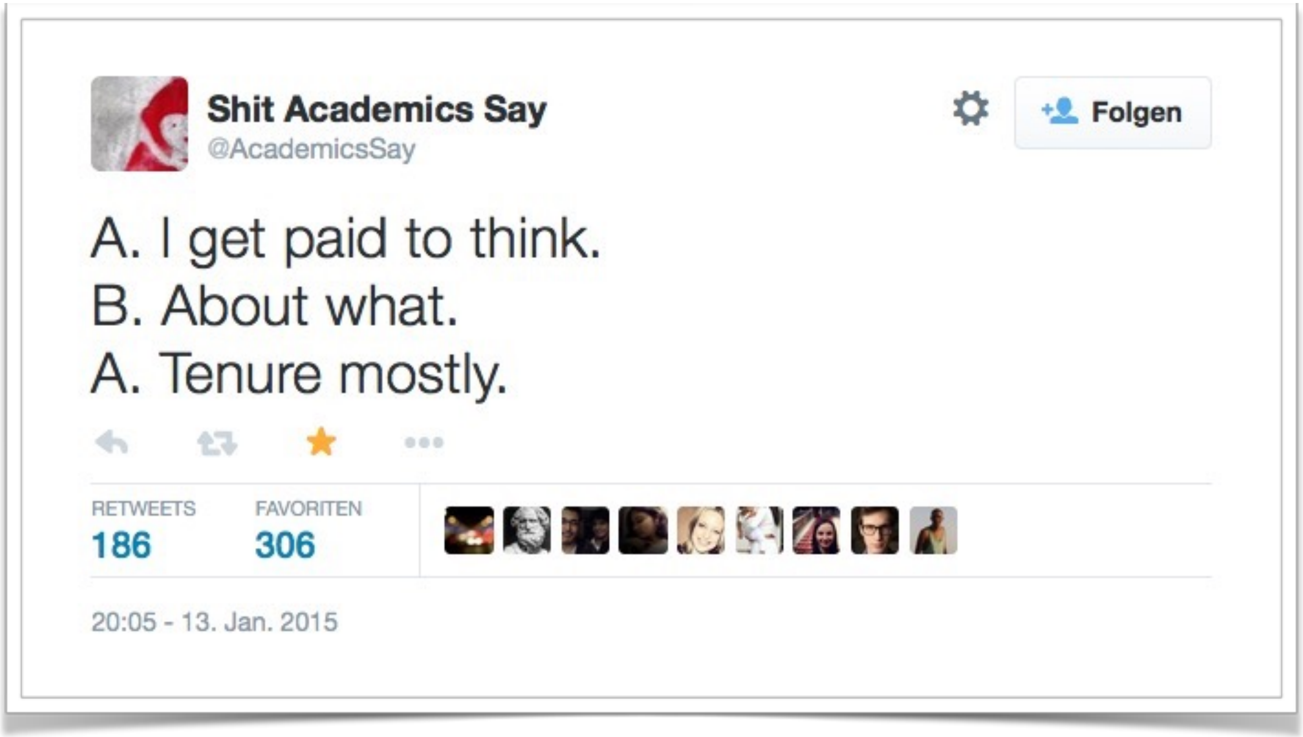
RESEARCH TRANSPARENCY

OSC
LMU Open Science Center

www.nicebread.de
www.researchtransparency.org
@nicebread303

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

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



Shit Academics Say
@AcademicsSay

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

RETWEETS 186
FAVORITEN 306

20:05 - 13. Jan. 2015

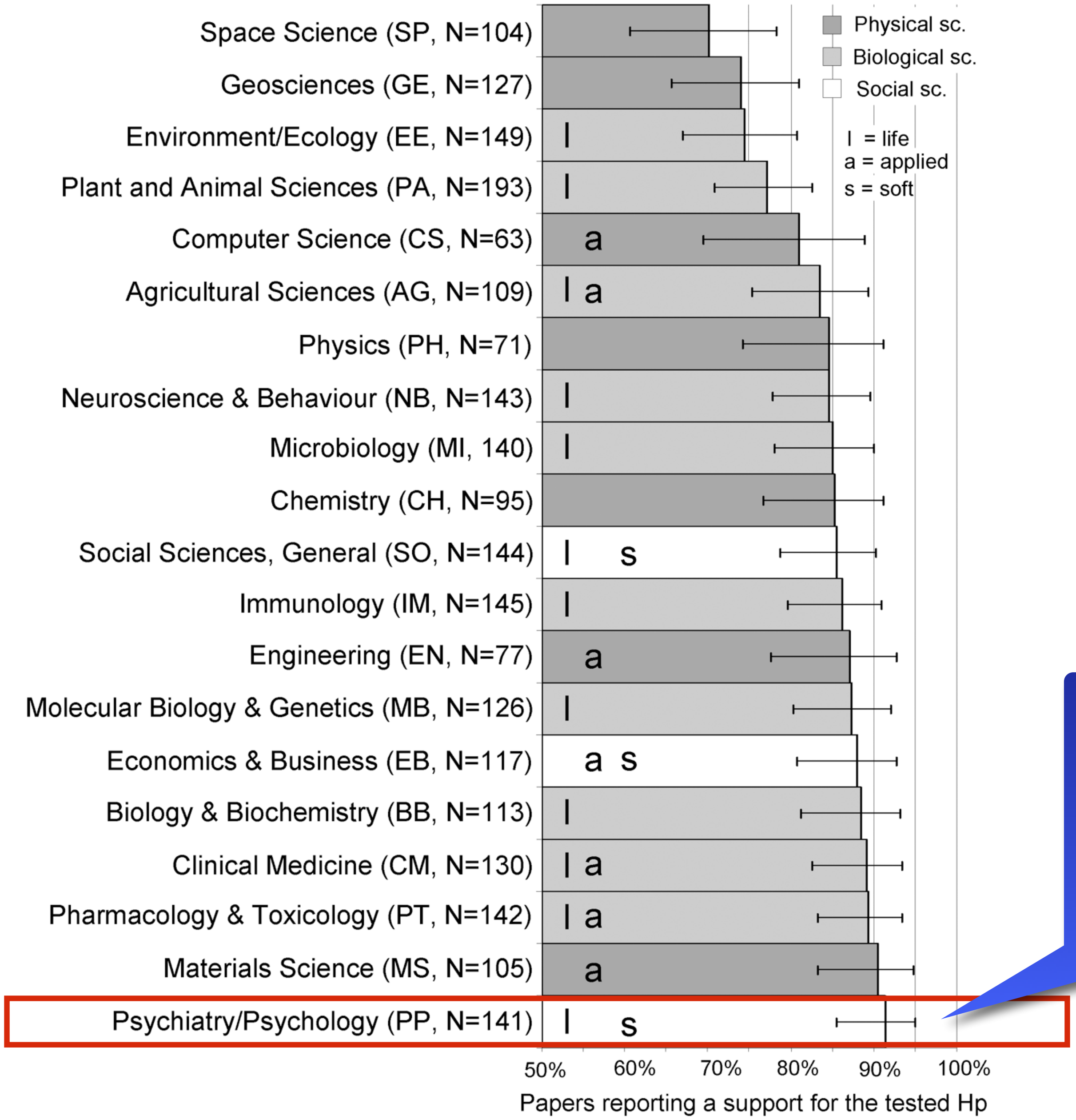
How to become a Professor?

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.

Abele-Brehm, A. E., & Bühner, M. (2016). Wer soll die Professur bekommen? Psychologische Rundschau, 67(4), 250–261. http://doi.org/10.1026/0033-3042/a000335

How to get lots of publications?



92% of published papers have significant, positive results

Fanelli, D. (2010). “Positive” Results Increase Down the Hierarchy of the Sciences. *PLOS ONE*, *5*, e10068. doi:10.1371/journal.pone.0010068

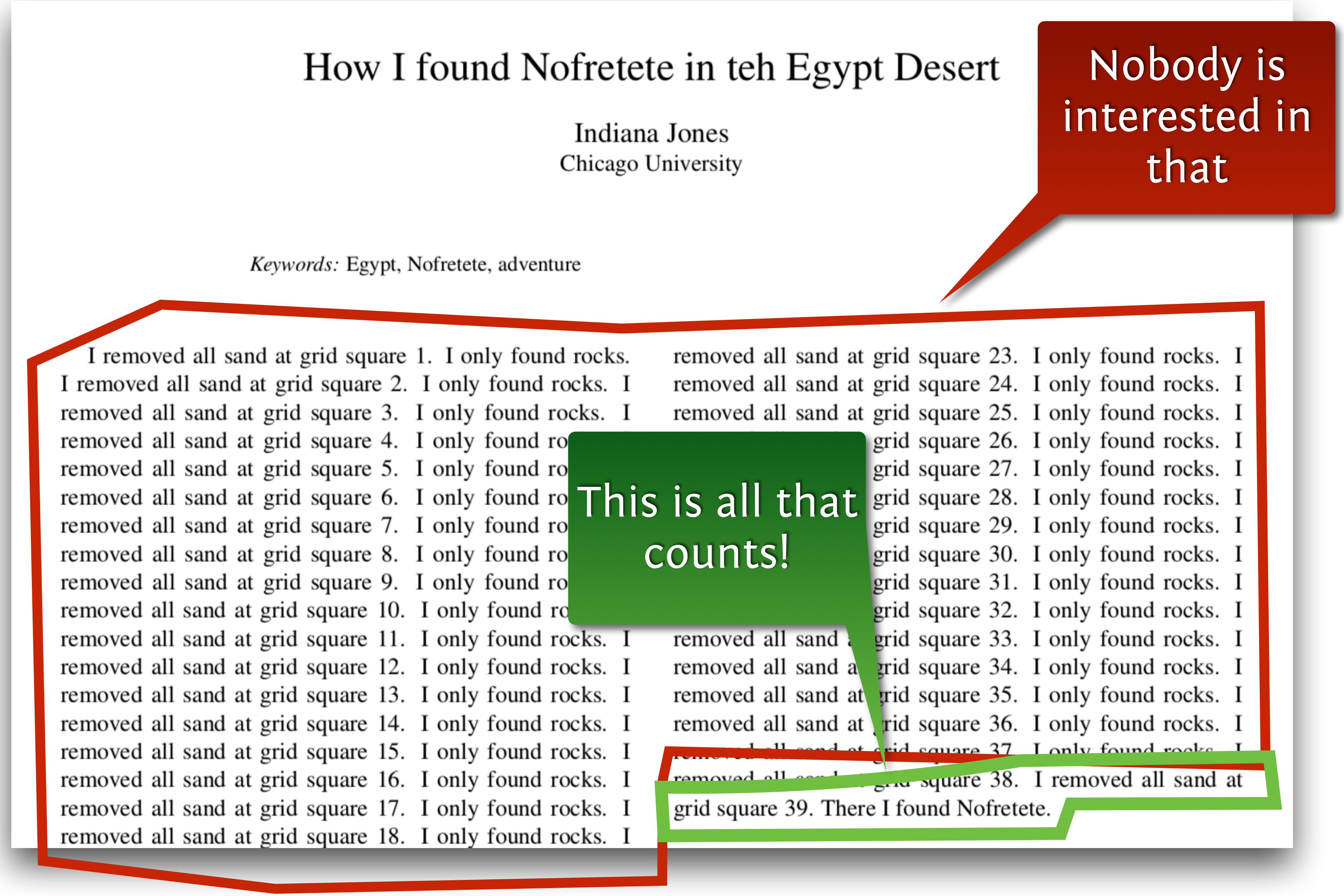
p-hack your way to scientific glory!

DOING RESEARCH WITH THE MINDSET OF AN ARCHAEOLOGIST

h-index = 241

DOING RESEARCH WITH THE MINDSET OF AN ARCHAEOLOGIST

h-index = 241



How I found Nofretete in teh Egypt Desert

Indiana Jones
Chicago University

Keywords: Egypt, Nofretete, adventure

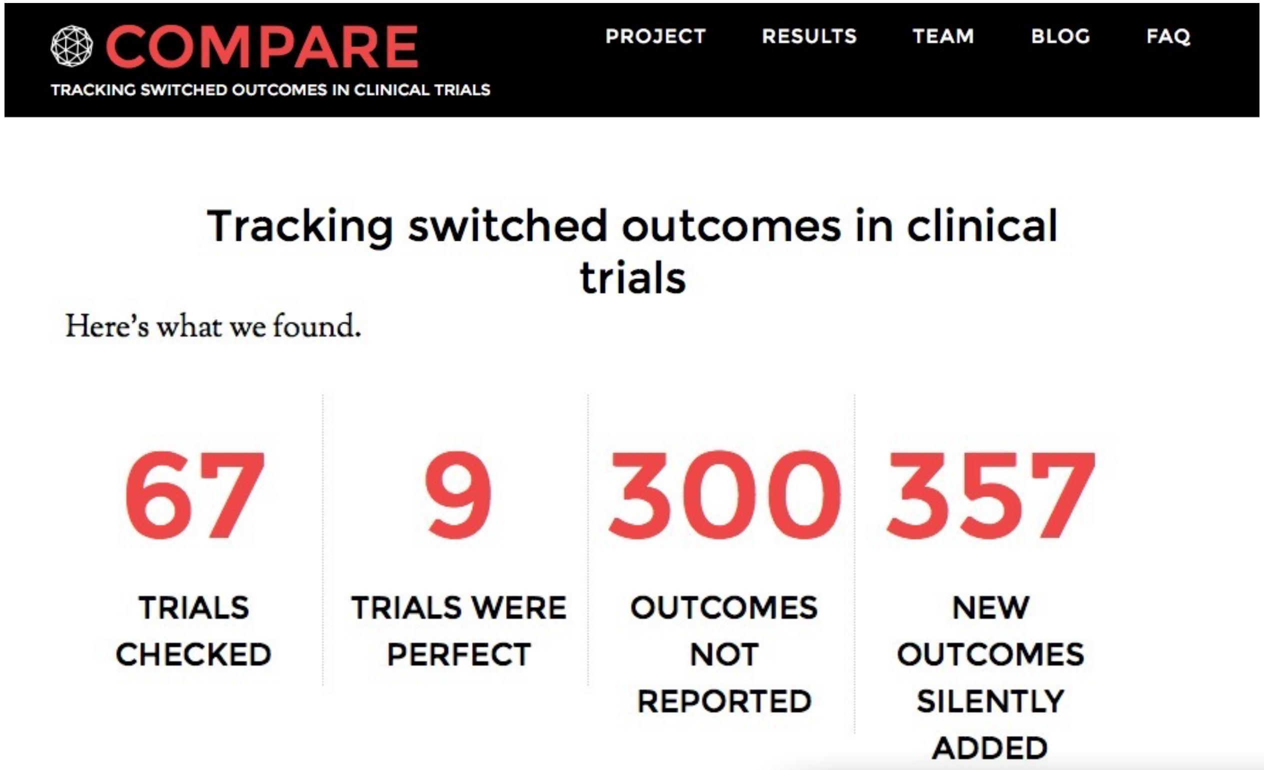
I removed all sand at grid square 1. I only found rocks.
I removed all sand at grid square 2. I only found rocks. I
removed all sand at grid square 3. I only found rocks. I
removed all sand at grid square 4. I only found ro
removed all sand at grid square 5. I only found ro
removed all sand at grid square 6. I only found ro
removed all sand at grid square 7. I only found ro
removed all sand at grid square 8. I only found ro
removed all sand at grid square 9. I only found ro
removed all sand at grid square 10. I only found ro
removed all sand at grid square 11. I only found rocks. I
removed all sand at grid square 12. I only found rocks. I
removed all sand at grid square 13. I only found rocks. I
removed all sand at grid square 14. I only found rocks. I
removed all sand at grid square 15. I only found rocks. I
removed all sand at grid square 16. I only found rocks. I
removed all sand at grid square 17. I only found rocks. I
removed all sand at grid square 18. I only found rocks. I

removed all sand at grid square 23. I only found rocks. I
removed all sand at grid square 24. I only found rocks. I
removed all sand at grid square 25. I only found rocks. I
grid square 26. I only found rocks. I
grid square 27. I only found rocks. I
grid square 28. I only found rocks. I
grid square 29. I only found rocks. I
grid square 30. I only found rocks. I
grid square 31. I only found rocks. I
grid square 32. I only found rocks. I
removed all sand grid square 33. I only found rocks. I
removed all sand a grid square 34. I only found rocks. I
removed all sand at grid square 35. I only found rocks. I
removed all sand at grid square 36. I only found rocks. I
removed all sand at grid square 37. I only found rocks. I
removed all sand at grid square 38. I removed all sand at
grid square 39. There I found Nofretete.

Nobody is interested in that

This is all that counts!

Tool 1: Outcome switching



COMPARE

TRACKING SWITCHED OUTCOMES IN CLINICAL TRIALS

PROJECT RESULTS TEAM BLOG FAQ

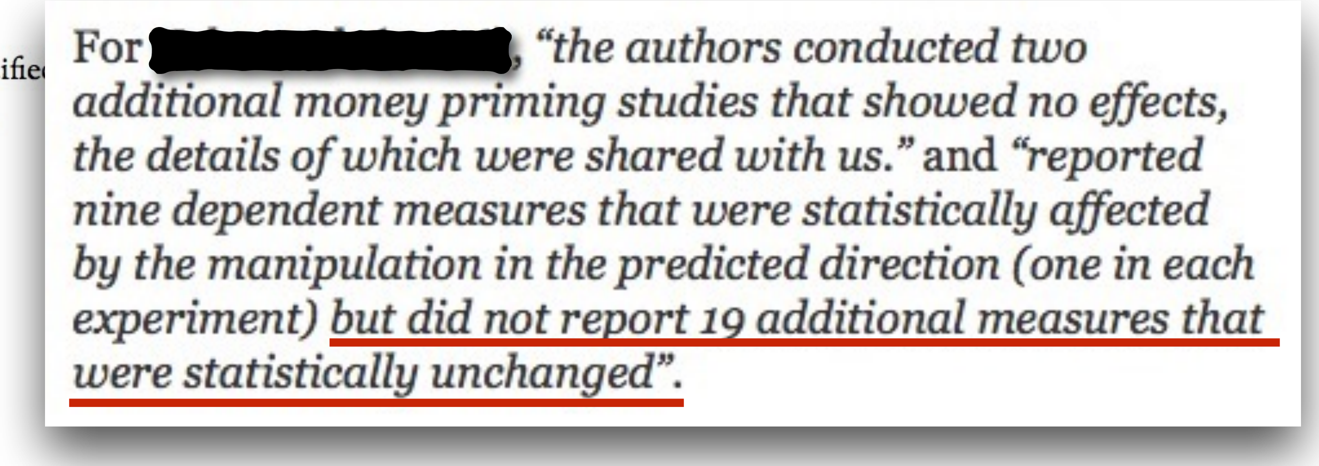
Tracking switched outcomes in clinical trials

Here's what we found.

67	9	300	357
TRIALS CHECKED	TRIALS WERE PERFECT	OUTCOMES NOT REPORTED	NEW OUTCOMES SILENTLY ADDED

On average, each trial reported just 62.1% of its specifie

average, each trial silently added 5.3 new outcomes.



For [redacted], *"the authors conducted two additional money priming studies that showed no effects, the details of which were shared with us."* and *"reported nine dependent measures that were statistically affected by the manipulation in the predicted direction (one in each experiment) but did not report 19 additional measures that were statistically unchanged".*

http://compare-trials.org/ http://blogs.discovermagazine.com/neuroskeptic/2015/07/23/social-priming-money-for-nothing/#.VuKRSRi5KJM

Tool 1: Outcome switching

- 2 outcome variables:

  false positive rate **5% → 9.5%**

- 5 outcome variables with one-sided testing:

  false positive rate **5% → 41%**

- How prevalent is it?
  - John, Loewenstein and Prelec (2012):
    66% of researchers admit having done this.

Tool 2: Many conditions, report only those that worked

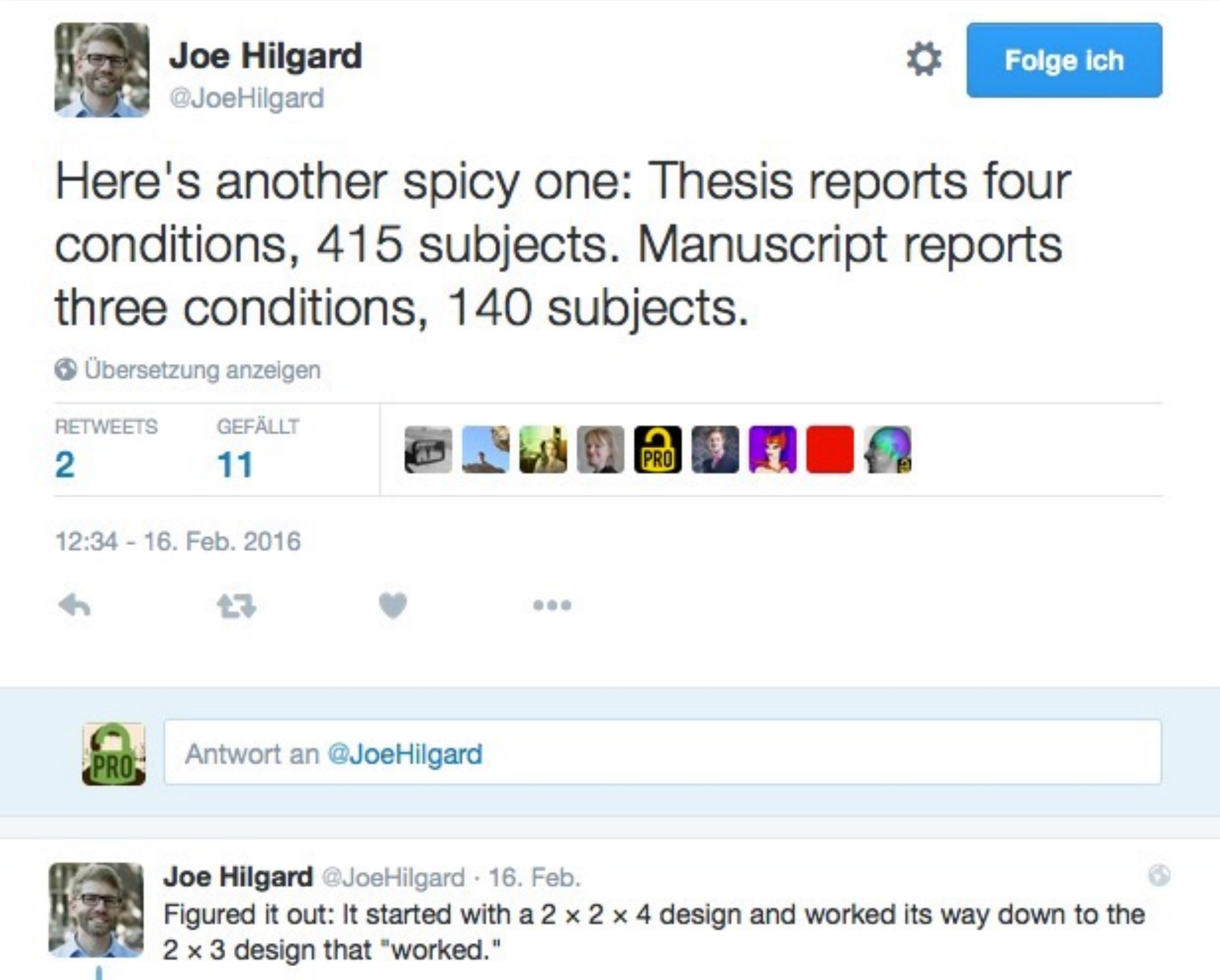
- Assess more than two conditions (and leave out conditions that are not significantly different).
- E.g., testing "high", "medium" and "low" conditions and reporting only the results of a "high" versus "medium" comparison.
- Gives you more than one chance to find an effect. Can increases the false positive rate to **12.6%**.
- How prevalent is it?
  - 27% of researchers admit having done this (John et al., 2012).

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22, 1359–1366. doi:10.1177/0956797611417632

Tool 2: Many conditions, report only those that worked

https://twitter.com/JoeHilgard/status/699693258386051072

Best-practice example: **Transform a boring dissertation into a groundbreaking publication (aka. „the Chrysalis Effect“; O’Boyle et al., 2014)**



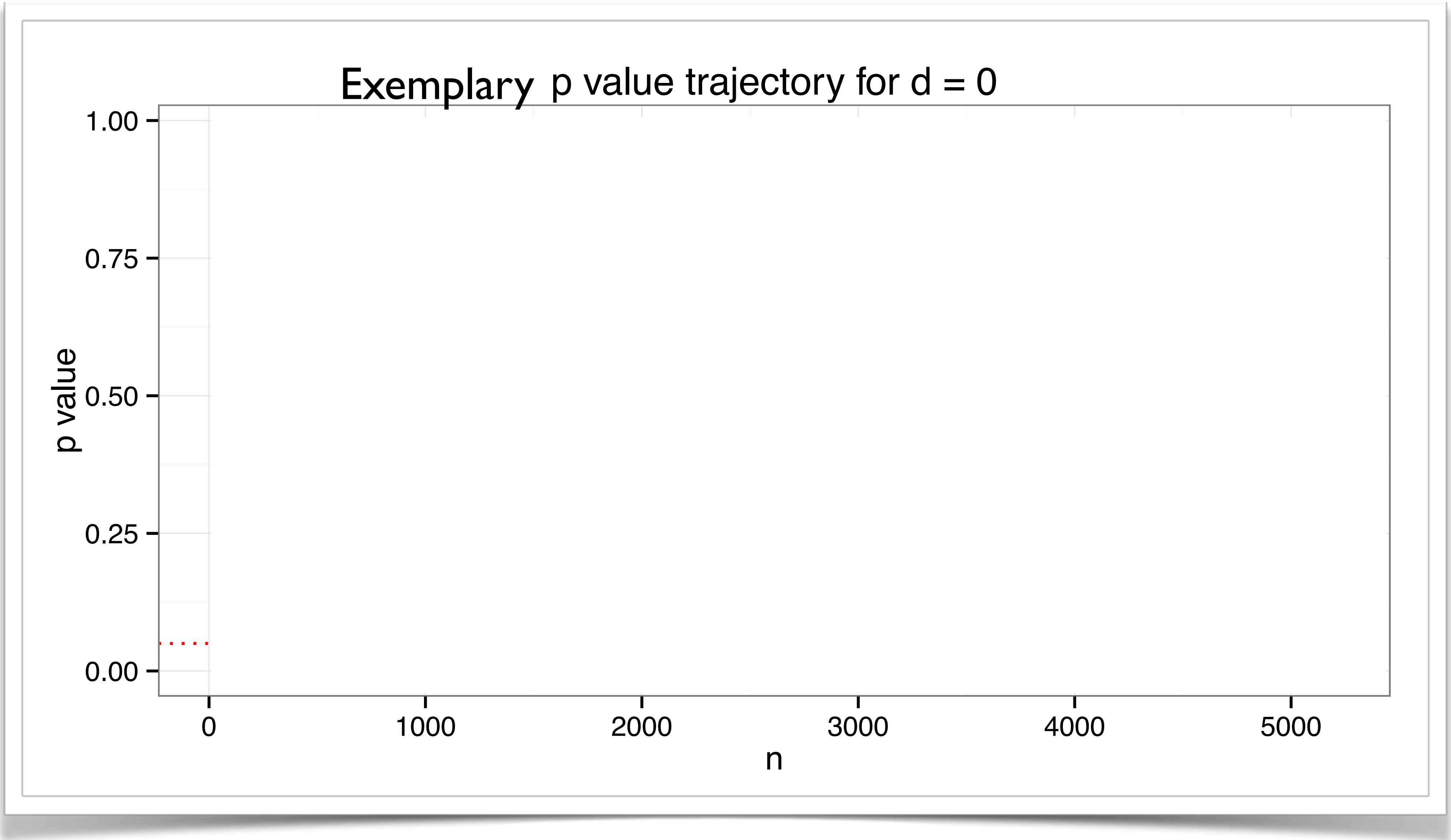
Joe Hilgard
@JoeHilgard

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

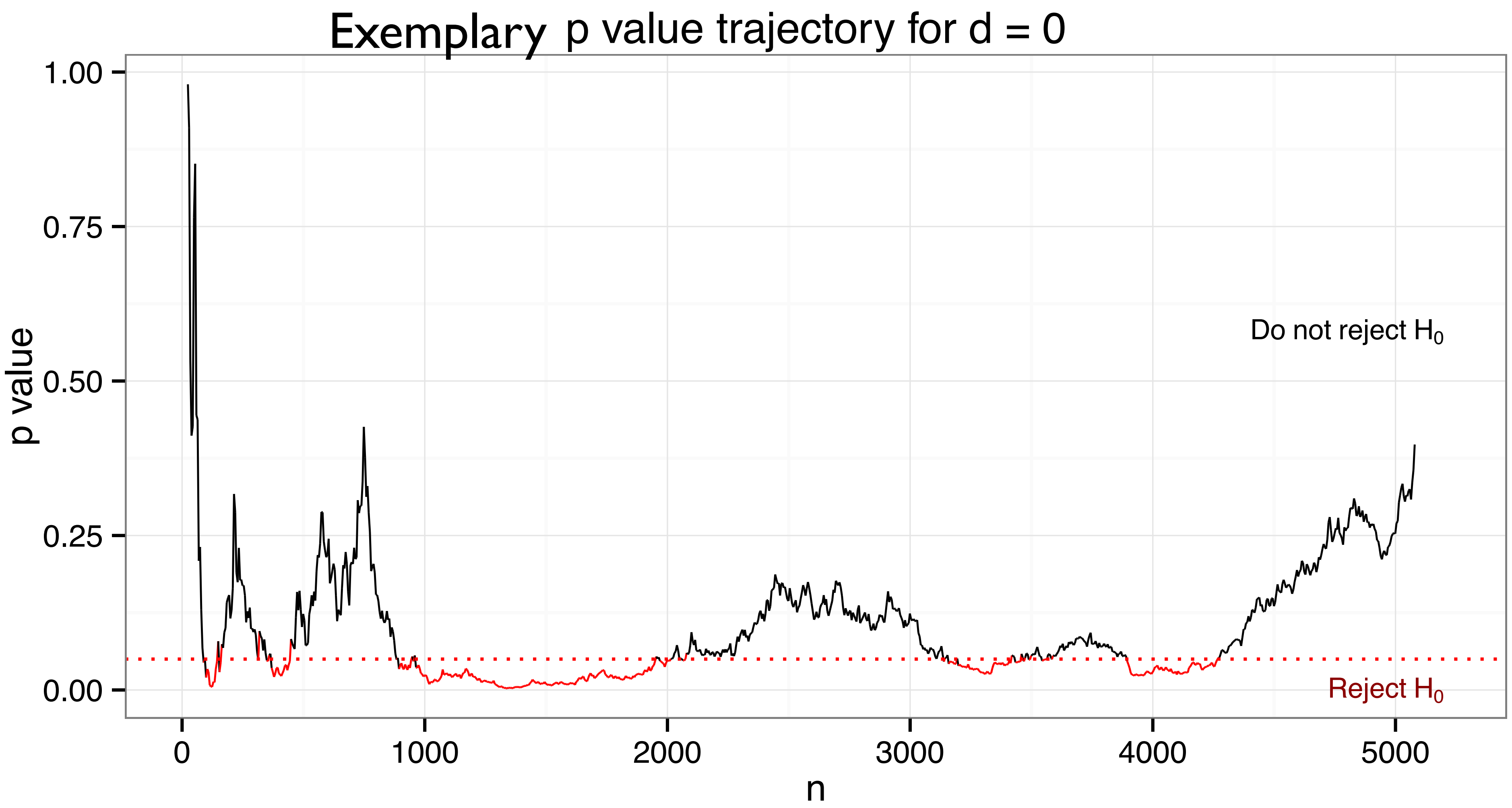
12:34 - 16. Feb. 2016

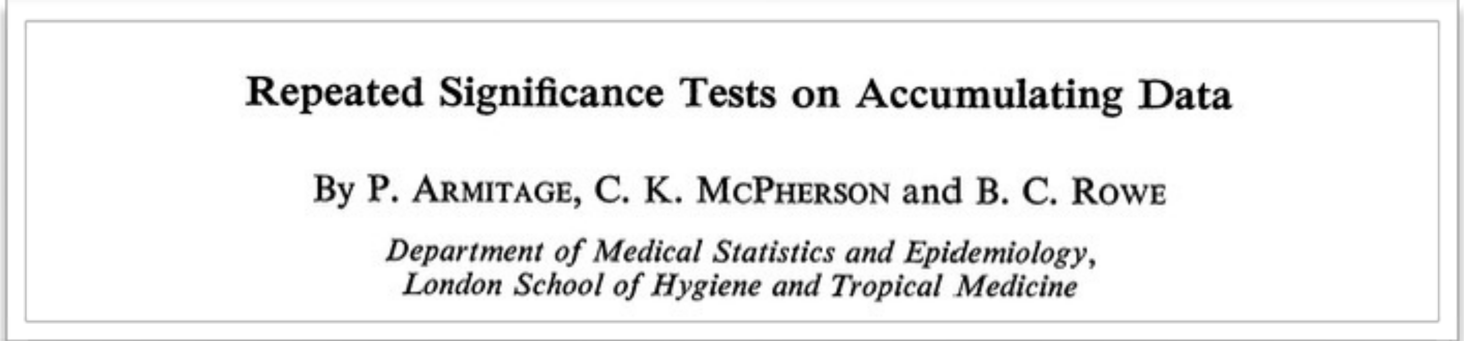
Joe Hilgard @JoeHilgard · 16. Feb.
Figured it out: It started with a 2 × 2 × 4 design and worked its way down to the 2 × 3 design that "worked."

Under H_0, *p* values meander infinitely



Under H_0, *p* values meander infinitely





Repeated Significance Tests on Accumulating Data

By P. Armitage, C. K. McPherson and B. C. Rowe

Department of Medical Statistics and Epidemiology,
London School of Hygiene and Tropical Medicine

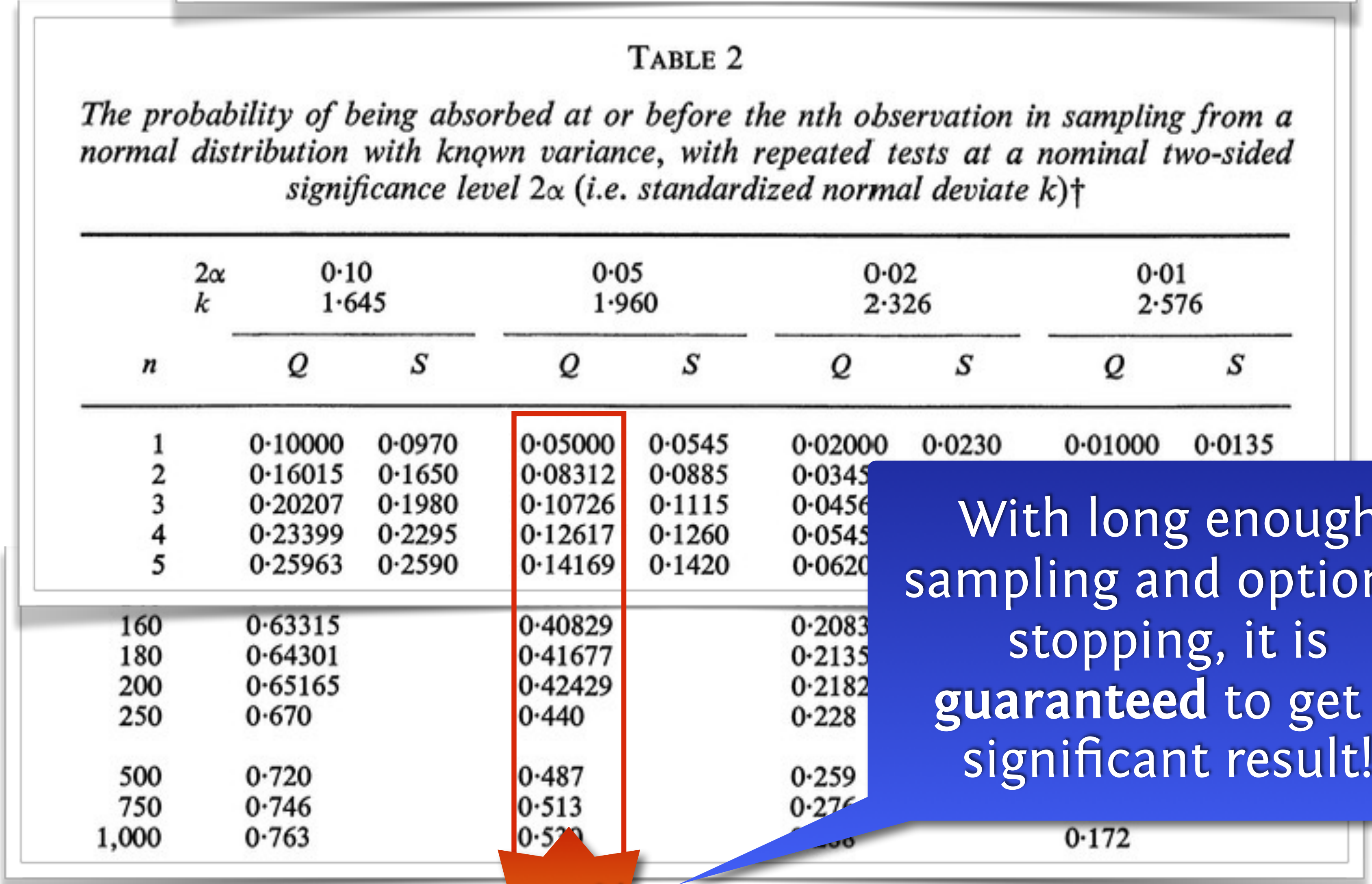


Table 2

The probability of being absorbed at or before the nth observation in sampling from a normal distribution with known variance, with repeated tests at a nominal two-sided significance level 2α (i.e. standardized normal deviate k)†

2α k	0·10 1·645		0·05 1·960		0·02 2·326		0·01 2·576	
n	*Q*	*S*	*Q*	*S*	*Q*	*S*	*Q*	*S*
1	0·10000	0·0970	0·05000	0·0545	0·02000	0·0230	0·01000	0·0135
2	0·16015	0·1650	0·08312	0·0885	0·0345			
3	0·20207	0·1980	0·10726	0·1115	0·0456			
4	0·23399	0·2295	0·12617	0·1260	0·0545			
5	0·25963	0·2590	0·14169	0·1420	0·0620			
160	0·63315		0·40829		0·2083			
180	0·64301		0·41677		0·2135			
200	0·65165		0·42429		0·2182			
250	0·670		0·440		0·228			
500	0·720		0·487		0·259			
750	0·746		0·513		0·27			
1,000	0·763		0·5		[illegible]		0·172	

100%

With long enough sampling and optional stopping, it is **guaranteed** to get a significant result!

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. Journal of the Royal Statistical Society. Series A (General), 132, 235–244.

Tool 3: Optional stopping

- Collect an initial sample, analyze the results, add additional participants if not significant, stop when significance is found
- Increase twice: α = **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).

Tool 4: Multiple comparisons in ANOVA

- ANOVA, 3 factors, full model
  - 3 main effects, 3 two-way interactions, 1 three-way interaction

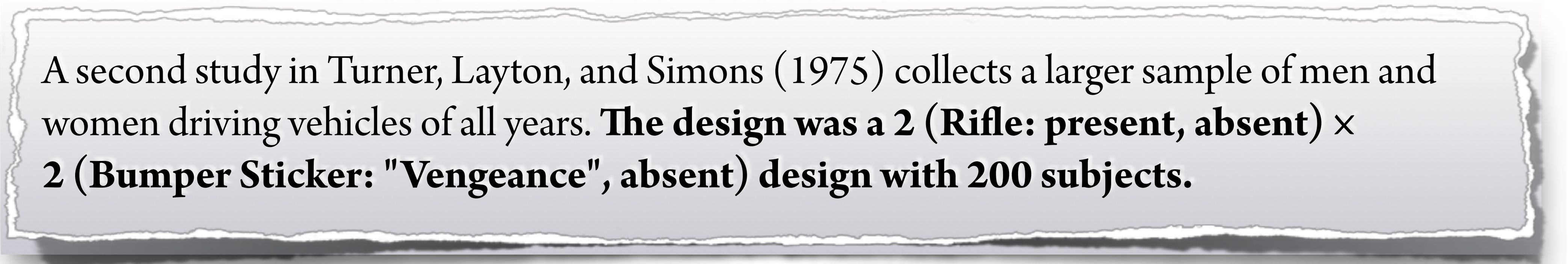
30.1%

- Type I error rate for at least 1 significant term?
- Well-Known: Corrections for post-hoc comparisons of levels within one factor
- Less-known: The need for correcting multiple interactions.

Cramer et al. (2013), Smith, Levine, & Lachlan (2002)

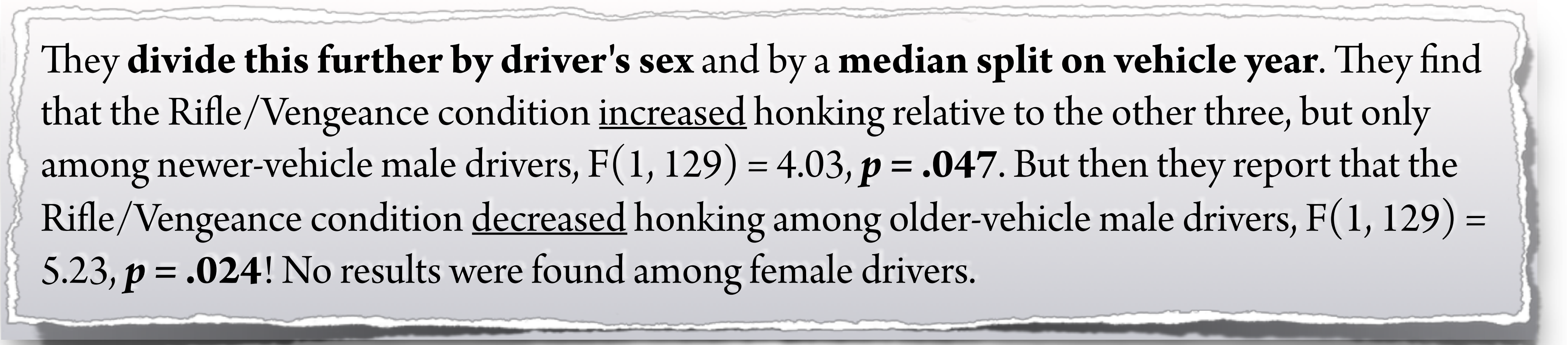
Tool 5: Subgroup analyses

Research question: Do aggressive primes 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.**

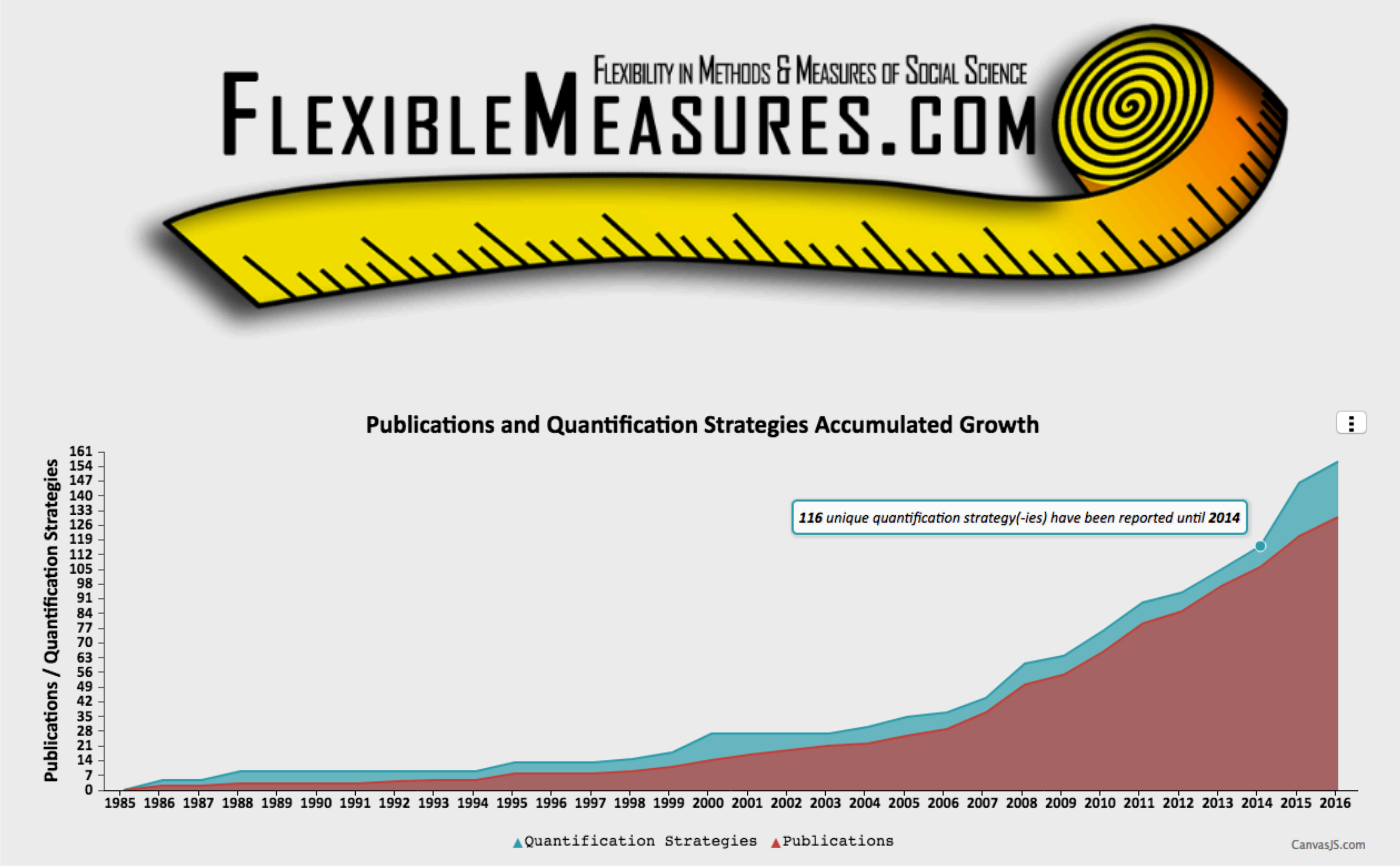
➙ presumably, no effect … (yet! Do not give up so easily)



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.

Cited from Joe Hilgard's excellent blog post on the Weapon Priming Effect: http://crystalprisonzone.blogspot.com/2016/03/the-weapons-priming-effect.html; Turner, C. W., Layton, J. F., & Simons, L. S. (1975). Naturalistic studies of aggressive behavior: aggressive stimuli, victim visibility, and horn honking. Journal of Personality and Social Psychology, 31(6), 1098–1107.

Tool 6: Flexible measures

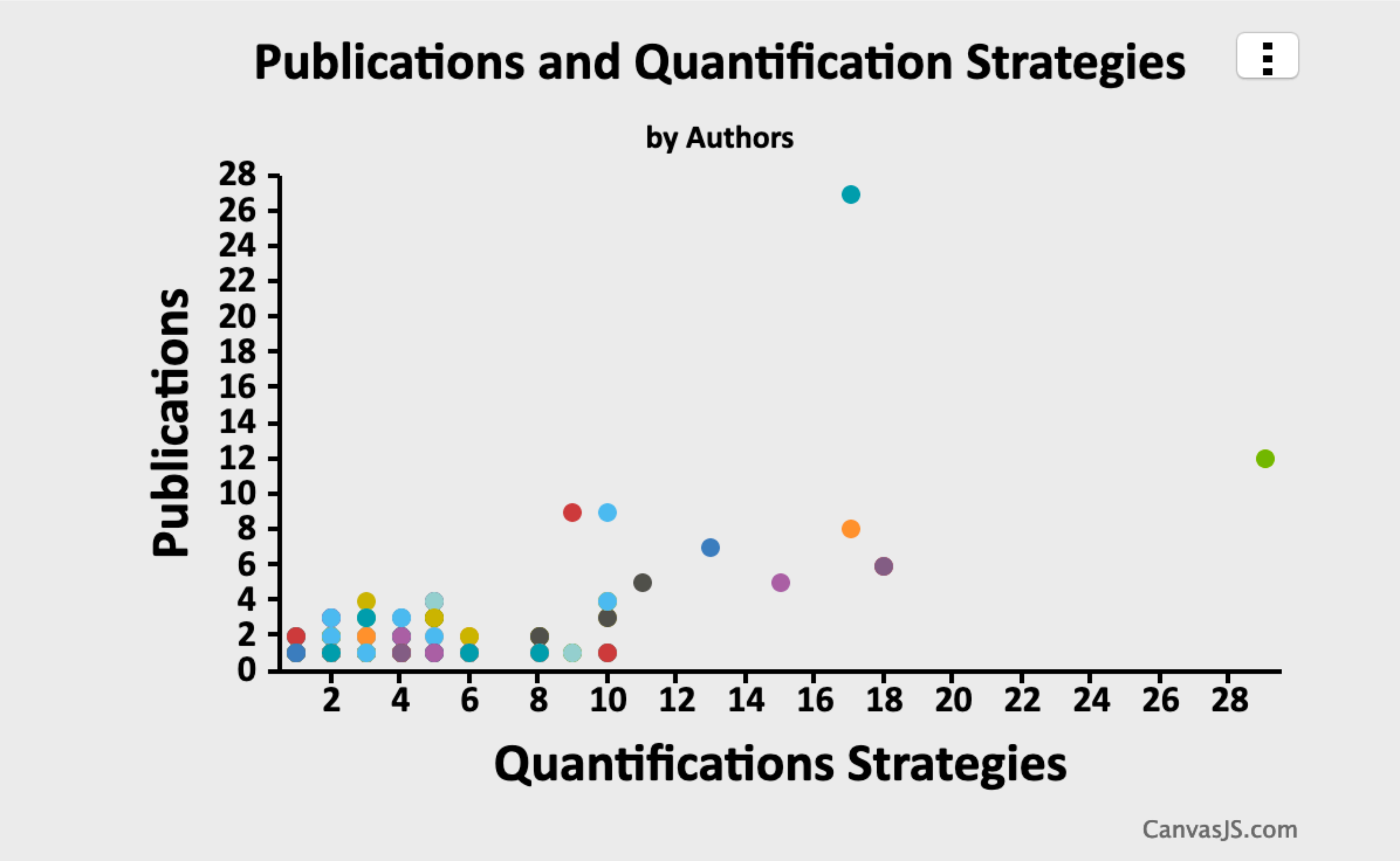


Publications and Quantification Strategies Accumulated Growth

*116 unique quantification strategy(-ies) have been reported until **2014***

http://www.flexiblemeasures.com/ by Malte Elson

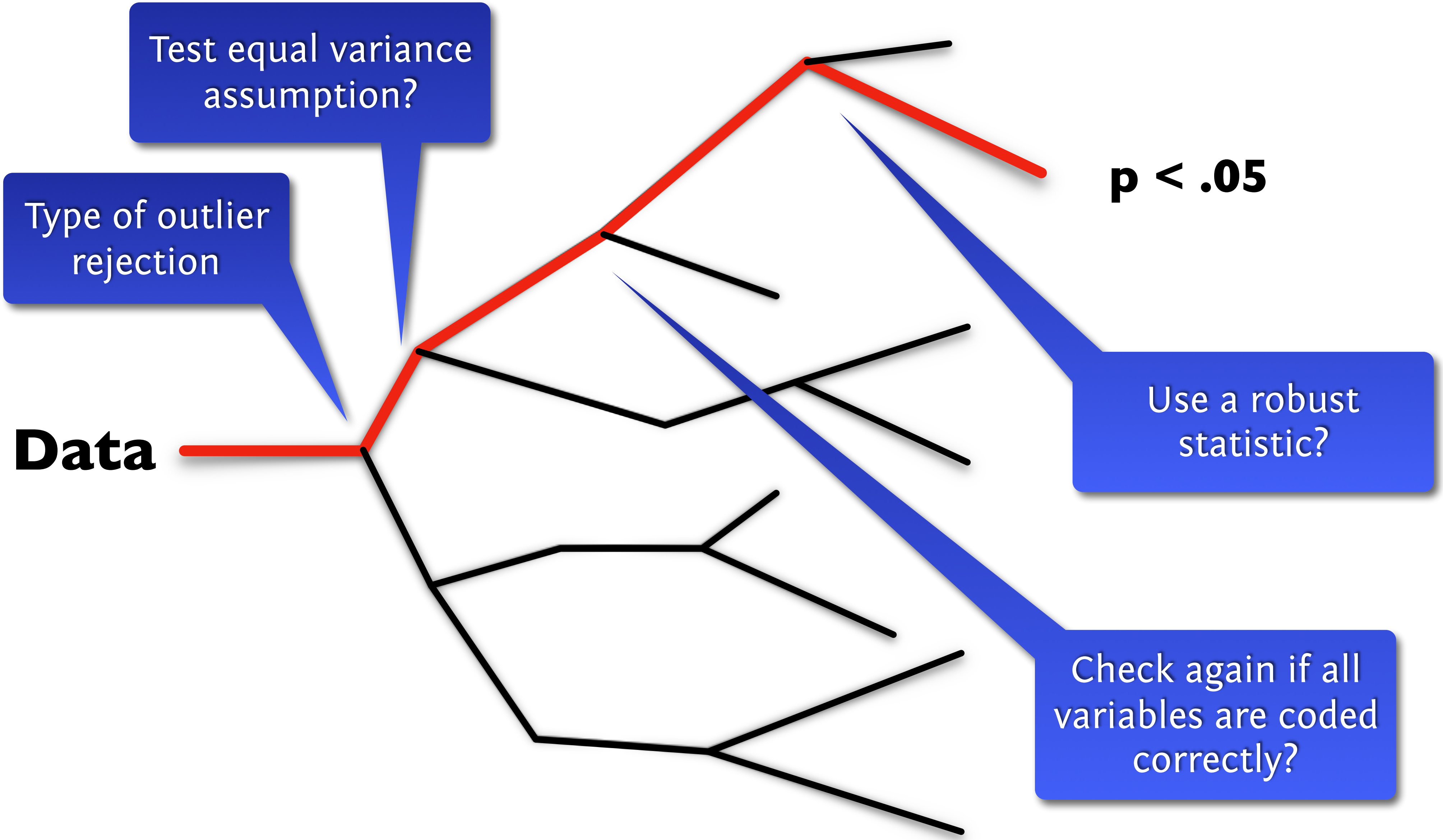
Tool 6: Flexible measures



http://www.flexiblemeasures.com/ by Malte Elson

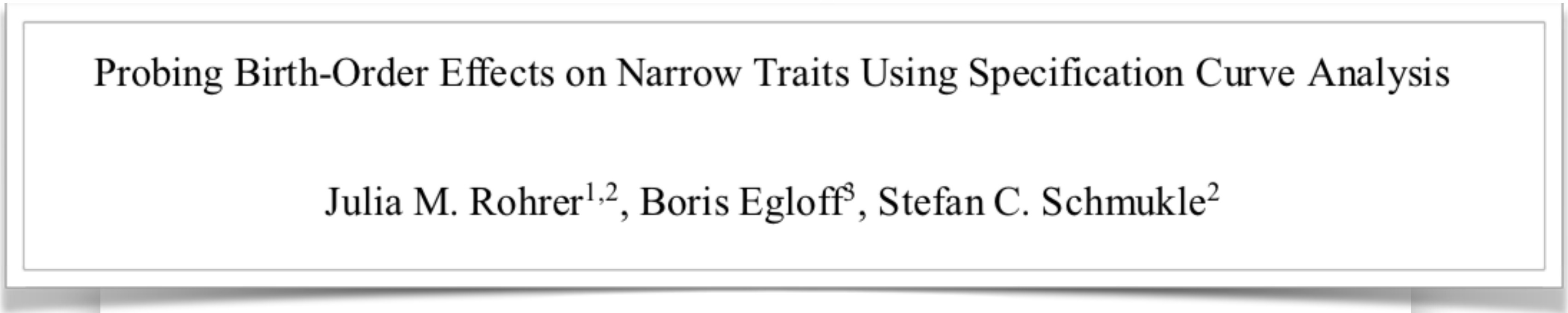
Tool 7: Explore the garden of forking paths

Andrew Gelman & Eric Loken, 2013



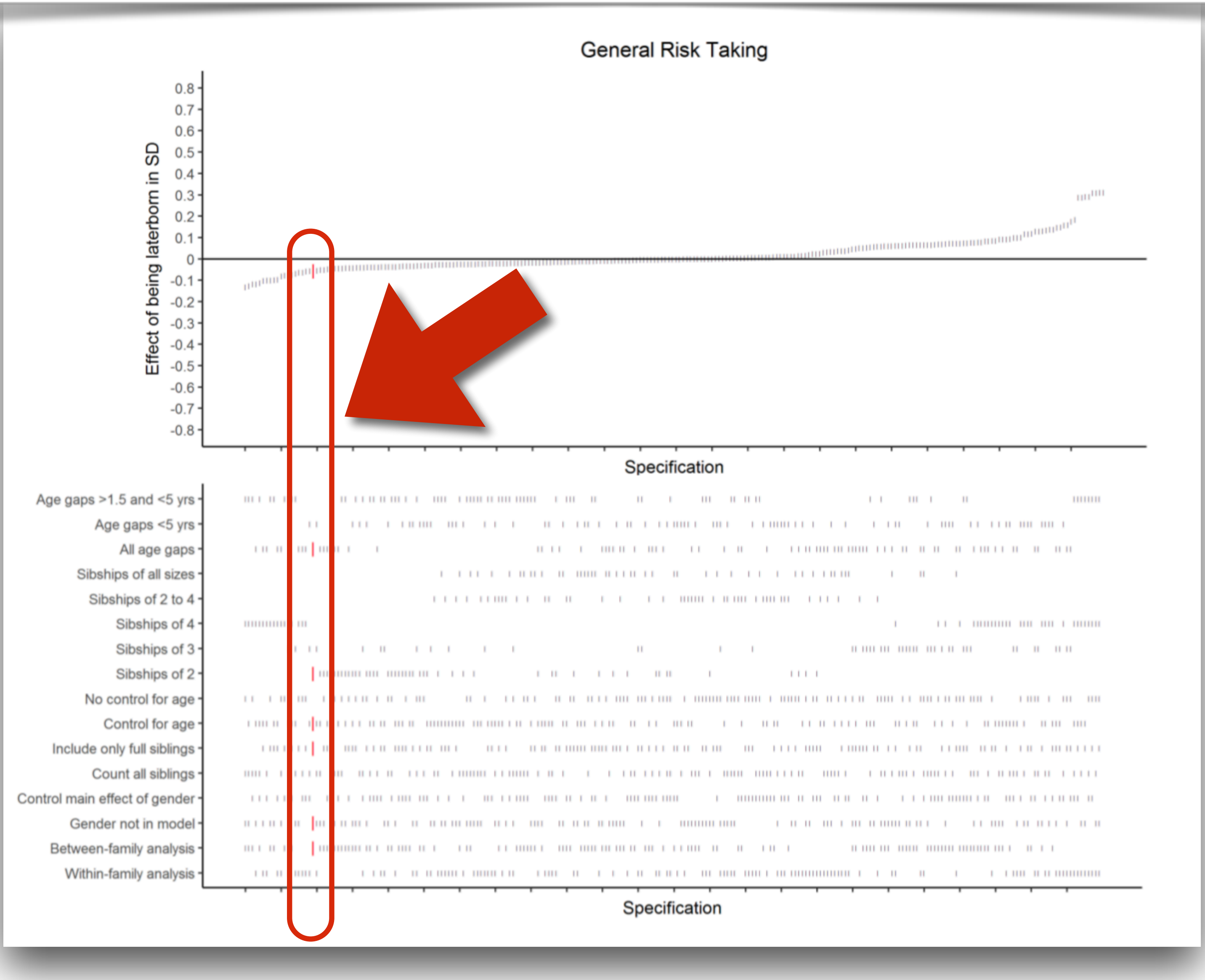
Test equal variance assumption?
Type of outlier rejection
Data
p < .05
Use a robust statistic?
Check again if all variables are coded correctly?

Inspired by Neurosceptic's blog: http://blogs.discovermagazine.com/neuroskeptic/2015/05/18/p-hacking-a-talk-and-further-thoughts/#.VV2TiOePKsN



Probing Birth-Order Effects on Narrow Traits Using Specification Curve Analysis

Julia M. Rohrer[1,2], Boris Egloff[3], Stefan C. Schmukle[2]



Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2017). Probing Birth-Order Effects on Narrow Traits Using Specification-Curve Analysis. *Psychological Science*, *28*, 1821–1832. doi:10.1177/0956797617723726

Tool 8: Build the *p*-hacking into the software!

CrossMark
click for updates

Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates

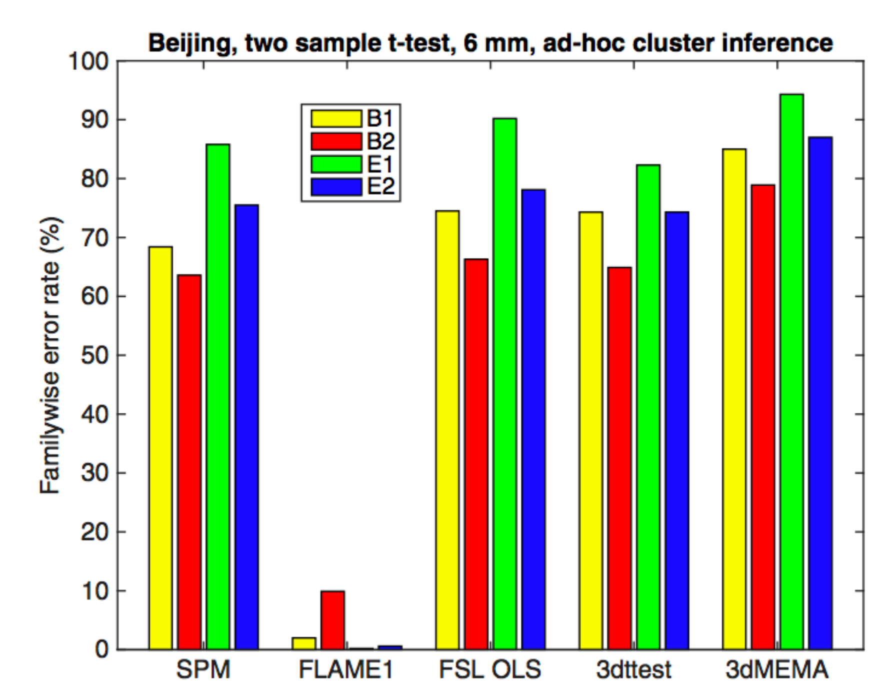
Anders Eklund[a,b,c,1], Thomas E. Nichols[d,e], and Hans Knutsson[a,c]

[a]Division of Medical Informatics, Department of Biomedical Engineering, Linköping University, S-581 85 Linköping, Sweden; [b]Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping Universi Science and Visualization, Linköping University, S-581 83 Linköping, Sweden; [d]Departmen Kingdom; and [e]WMG, University of Warwick, Coventry CV4 7AL, United Kingdom

Edited by Emery N. Brown, Massachusetts General Hospital, Boston, MA, and approved M

The most widely used task functional magnetic resonance imaging (fMRI) analyses use parametric statistical methods that depend on a variety of assumptions. In this work, we use real resting-state data and a total of 3 million random task group analyses to compute

(FWE), measure any sign one-sam



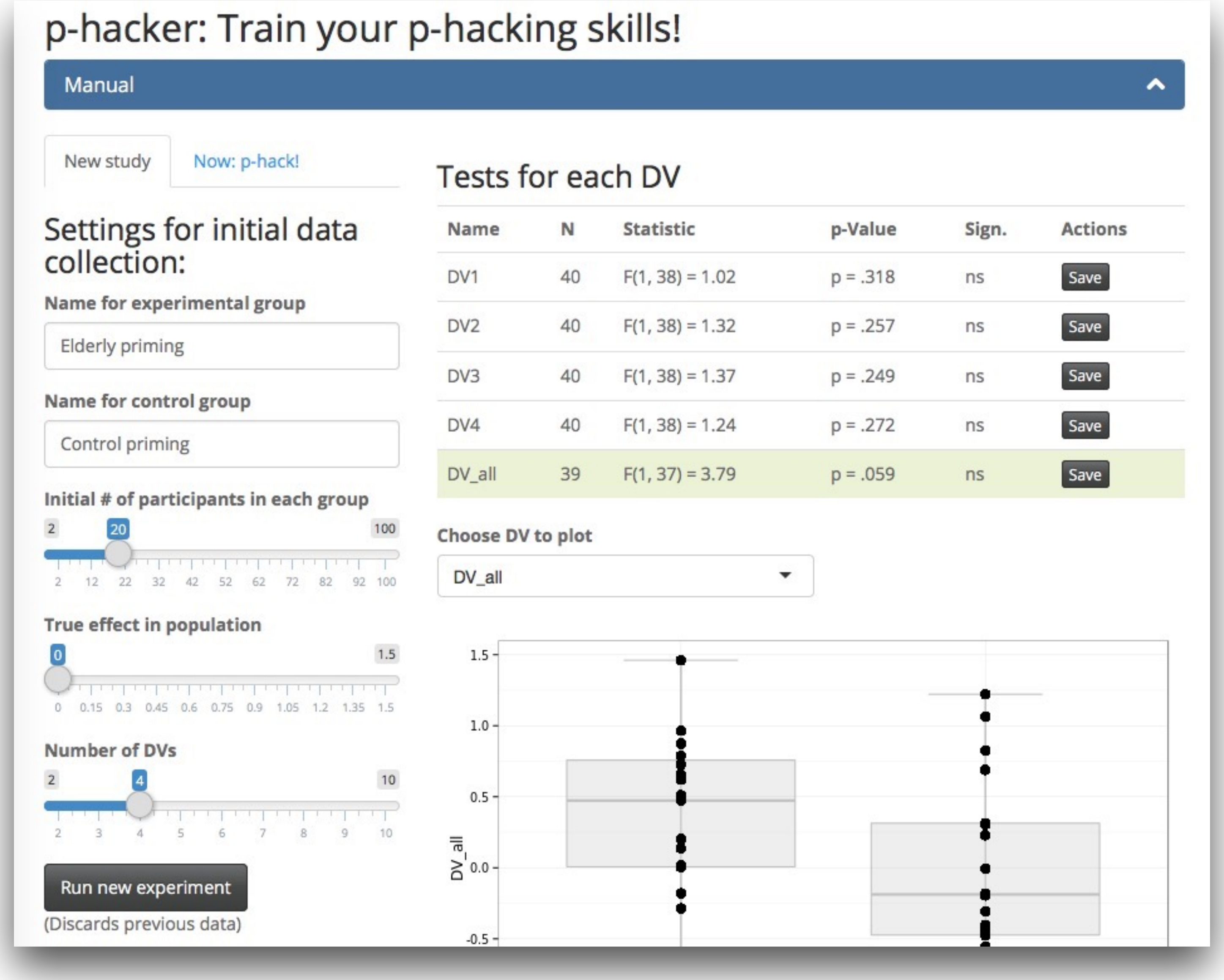
Beijing, two sample t-test, 6 mm, ad-hoc cluster inference

Freedom is nothing else but a chance to be better.

Albert Camus (1913 - 1960)

Ok, let’s celebrate some researcher degrees of freedom to be better at research!

Train your skills!



p-hacker: Train your p-hacking skills!

Settings for initial data collection:

Name for experimental group: Elderly priming

Name for control group: Control priming

Initial # of participants in each group: 20

True effect in population: 0

Number of DVs: 4

(Discards previous data)

Tests for each DV

Name	N	Statistic	p-Value	Sign.	Actions
DV1	40	F(1, 38) = 1.02	p = .318	ns	Save
DV2	40	F(1, 38) = 1.32	p = .257	ns	Save
DV3	40	F(1, 38) = 1.37	p = .249	ns	Save
DV4	40	F(1, 38) = 1.24	p = .272	ns	Save
DV_all	39	F(1, 37) = 3.79	p = .059	ns	Save

Choose DV to plot: DV_all

http://shinyapps.org/apps/p-hacker/

The impact of *p*-hacking on the rate of significant results

It is done …

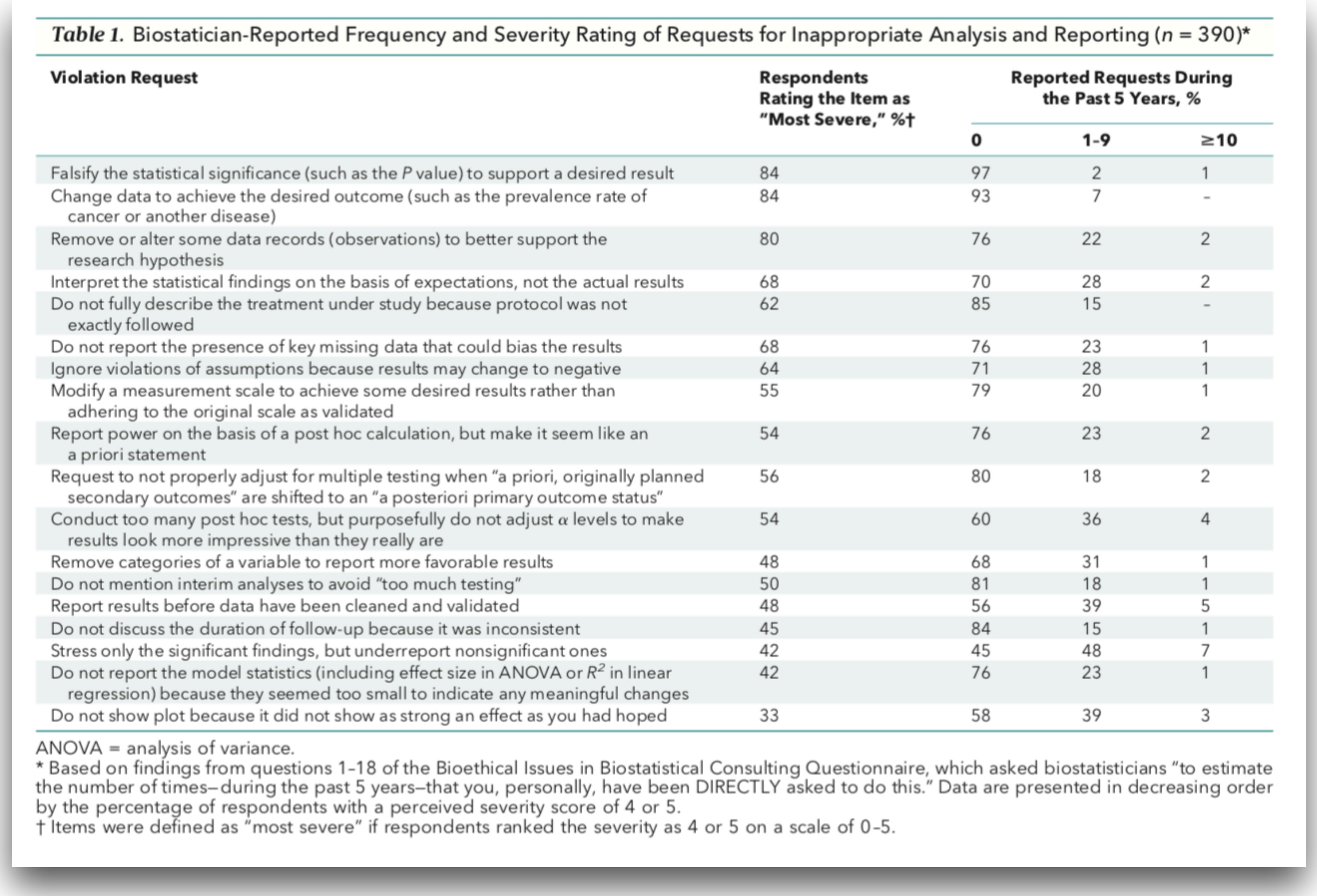


Table 1. Biostatician-Reported Frequency and Severity Rating of Requests for Inappropriate Analysis and Reporting (n = 390)*

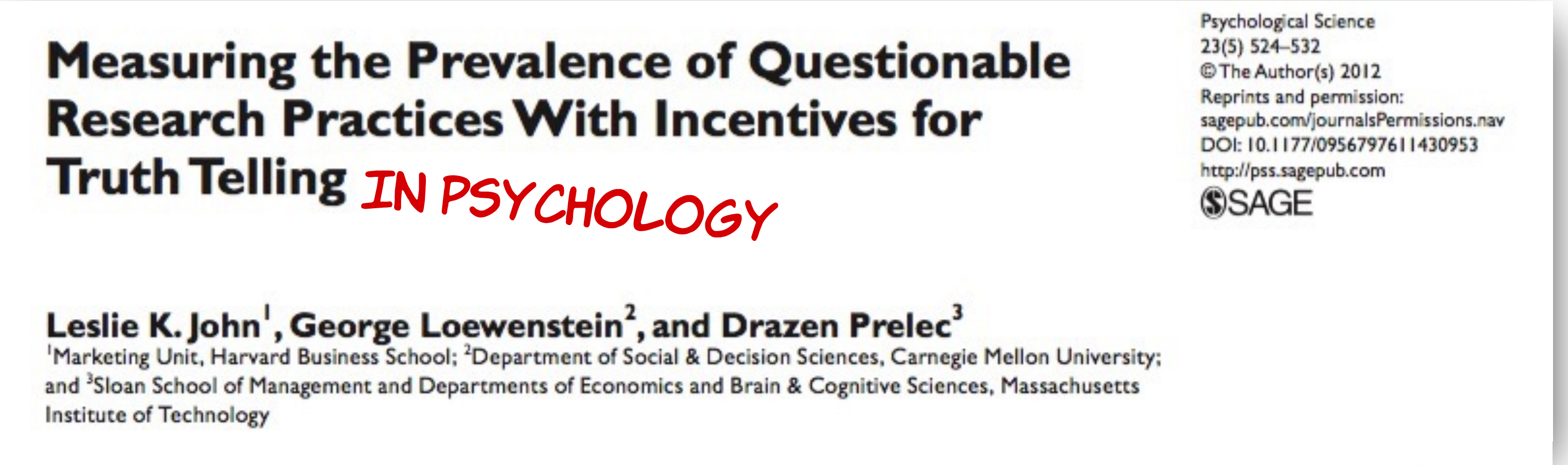
Violation Request	**Respondents Rating the Item as "Most Severe," %†**	**Reported Requests During the Past 5 Years, %**		
		0	**1-9**	**≥10**
Falsify the statistical significance (such as the *P* value) to support a desired result	84	97	2	1
Change data to achieve the desired outcome (such as the prevalence rate of cancer or another disease)	84	93	7	-
Remove or alter some data records (observations) to better support the research hypothesis	80	76	22	2
Interpret the statistical findings on the basis of expectations, not the actual results	68	70	28	2
Do not fully describe the treatment under study because protocol was not exactly followed	62	85	15	-
Do not report the presence of key missing data that could bias the results	68	76	23	1
Ignore violations of assumptions because results may change to negative	64	71	28	1
Modify a measurement scale to achieve some desired results rather than adhering to the original scale as validated	55	79	20	1
Report power on the basis of a post hoc calculation, but make it seem like an a priori statement	54	76	23	2
Request to not properly adjust for multiple testing when "a priori, originally planned secondary outcomes" are shifted to an "a posteriori primary outcome status"	56	80	18	2
Conduct too many post hoc tests, but purposefully do not adjust α levels to make results look more impressive than they really are	54	60	36	4
Remove categories of a variable to report more favorable results	48	68	31	1
Do not mention interim analyses to avoid "too much testing"	50	81	18	1
Report results before data have been cleaned and validated	48	56	39	5
Do not discuss the duration of follow-up because it was inconsistent	45	84	15	1
Stress only the significant findings, but underreport nonsignificant ones	42	45	48	7
Do not report the model statistics (including effect size in ANOVA or R^2 in linear regression) because they seemed too small to indicate any meaningful changes	42	76	23	1
Do not show plot because it did not show as strong an effect as you had hoped	33	58	39	3

ANOVA = analysis of variance.

* Based on findings from questions 1-18 of the Bioethical Issues in Biostatistical Consulting Questionnaire, which asked biostatisticians "to estimate the number of times–during the past 5 years–that you, personally, have been DIRECTLY asked to do this." Data are presented in decreasing order by the percentage of respondents with a perceived severity score of 4 or 5.

† Items were defined as "most severe" if respondents ranked the severity as 4 or 5 on a scale of 0-5.

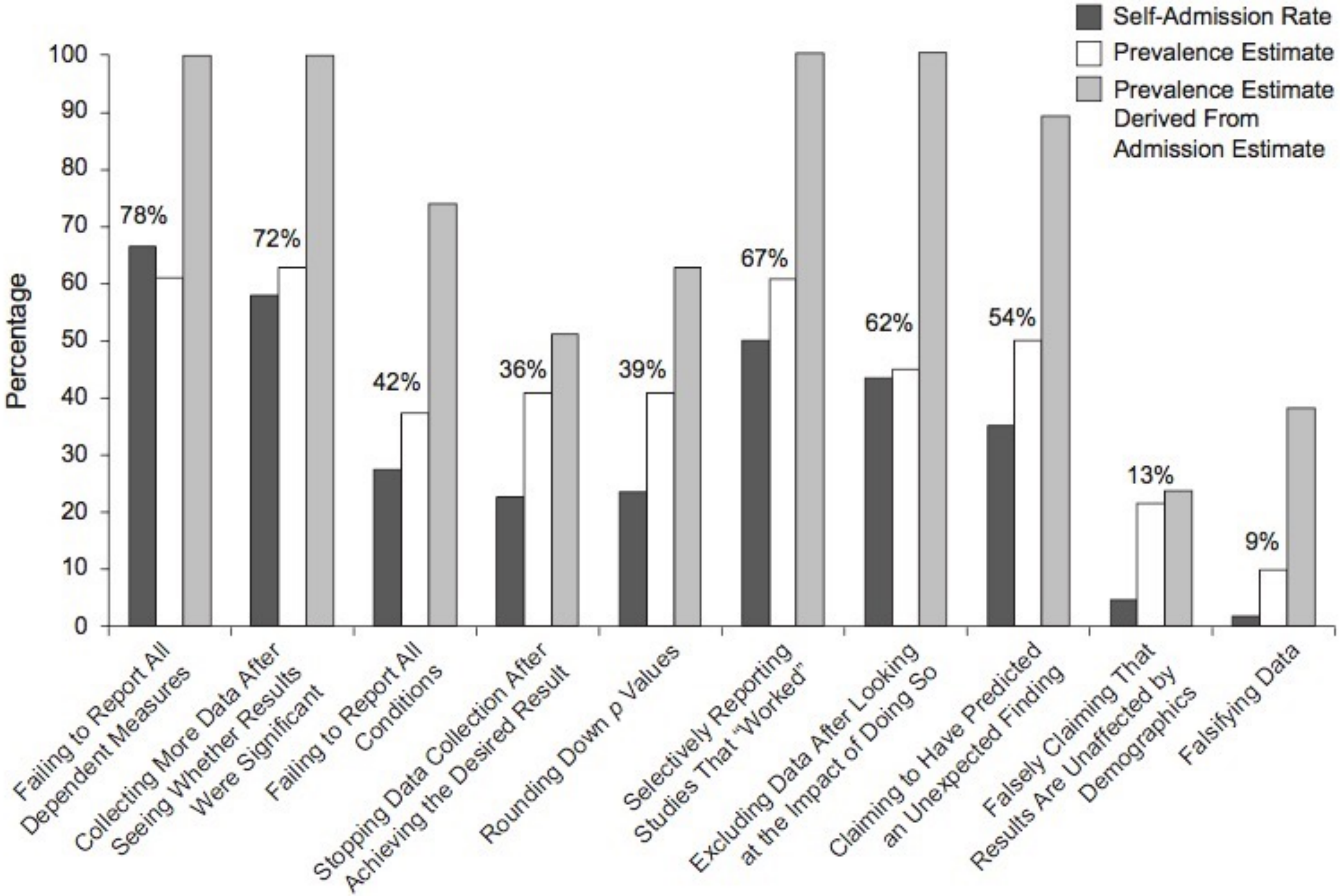
Wang M, Yan AF, & Katz RV. (2018). Researcher requests for inappropriate analysis and reporting: A u.s. survey of consulting biostatisticians. *Annals of Internal Medicine*. doi:10.7326/M18-1230



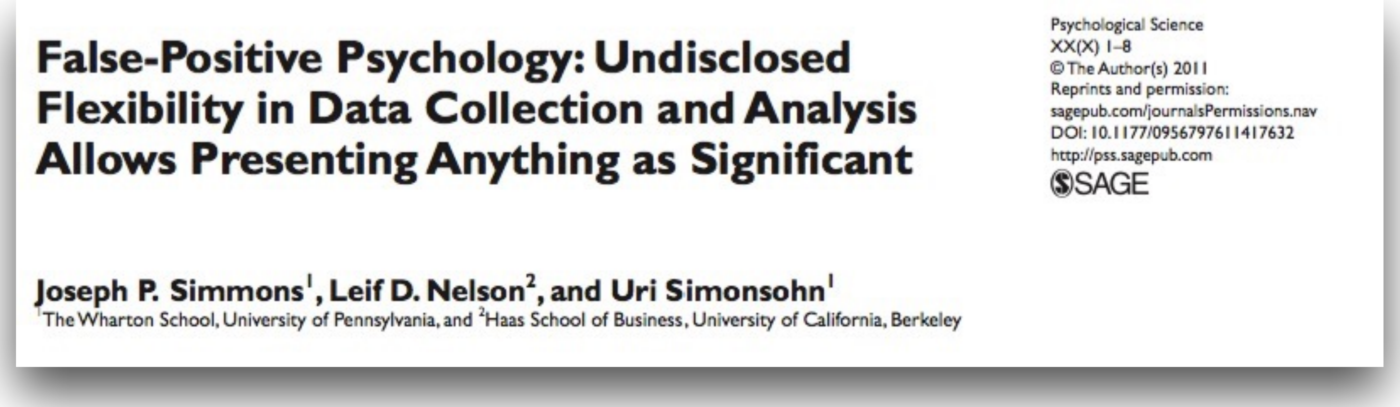
Psychological Science
23(5) 524–532
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611430953
http://pss.sagepub.com
SAGE

Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling IN PSYCHOLOGY

Leslie K. John[1]**, George Loewenstein**[2]**, and Drazen Prelec**[3]
[1]Marketing Unit, Harvard Business School; [2]Department of Social & Decision Sciences, Carnegie Mellon University; and [3]Sloan School of Management and Departments of Economics and Brain & Cognitive Sciences, Massachusetts Institute of Technology



How effective can it be?



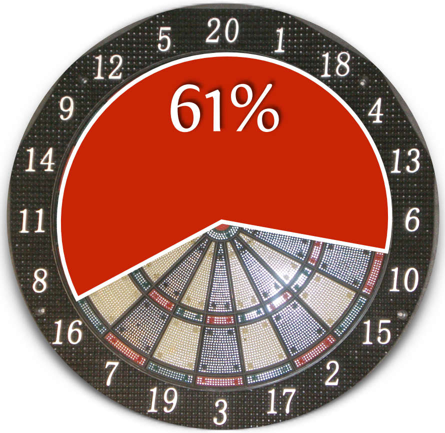
False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Psychological Science
XX(X) 1–8
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]

[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

- Doing some of these “questionable research practices” (QRPs) in combination raises the rate of significant results under H_0 from 5% to **<u>61%</u>**!



Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22, 1359–1366. doi:10.1177/0956797611417632

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. American Statistician, 00–00. http://doi.org/10.1080/00031305.2016.1154108

How effective can it be?

- From a statistical point of view, *p*-hacking increases your statistical power

$$Pr(p < .05 \mid H_1, phack) > Pr(p < .05 \mid H_1)$$

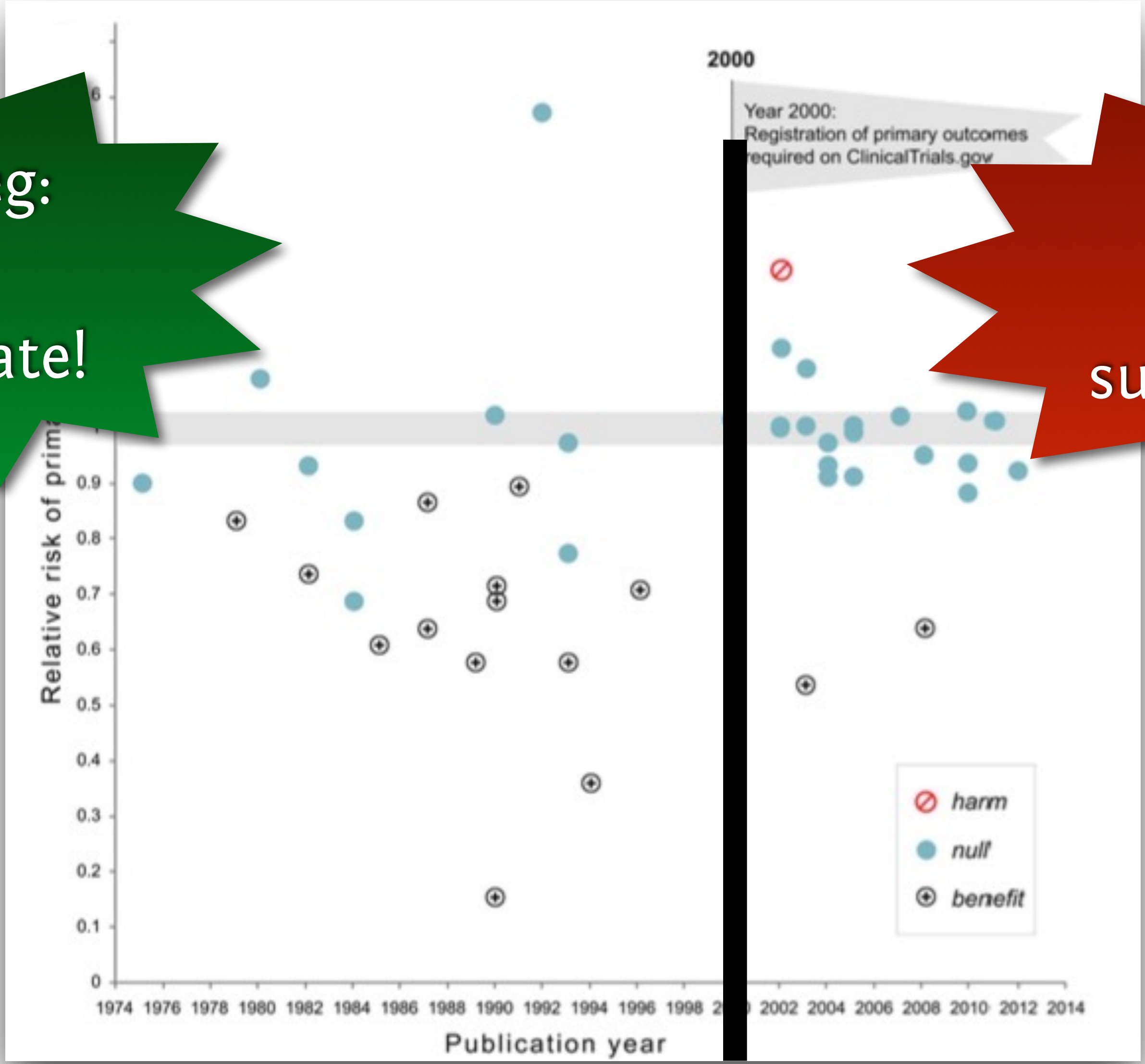
- For example:
  - Meta-analysis with *k* = 10 studies, true effect is **δ** = 0.2, typical sample sizes
  - Power **without** *p*-hacking in primary studies: **53%**
  - Power **with** *p*-hacking in primary studies: **76%**

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31234/osf.io/9h3nu

Things to avoid

Anti-tool:
Pre-registration stops *p*-hacking

DANGER



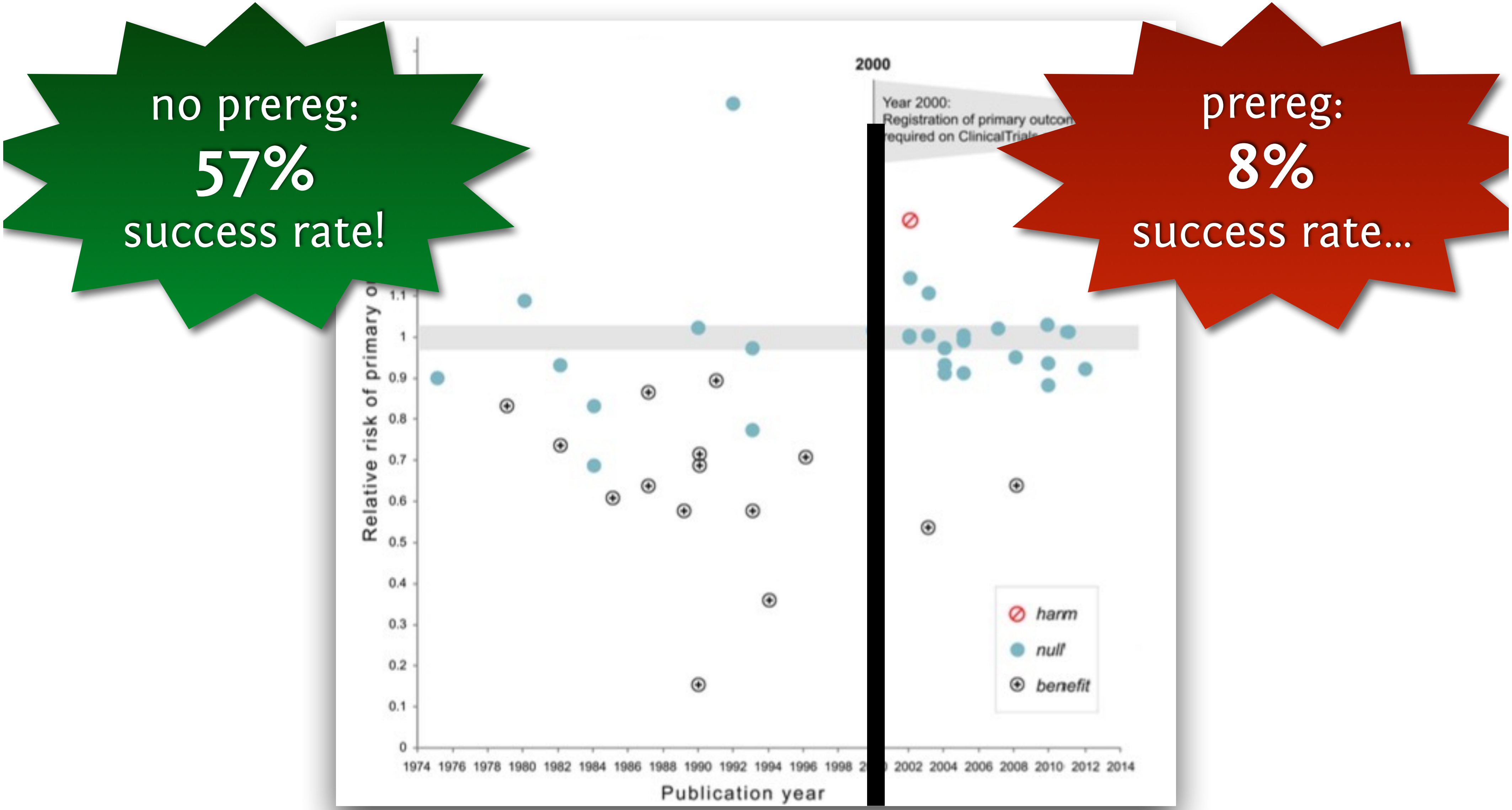
no prereg:
57%
success rate!

prereg:
8%
success rate...

http://chrisblattman.com/2016/03/01/13719/

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLoS ONE*, *10*(8), e0132382–12. http://doi.org/10.1371/journal.pone.0132382

Tool 9: Do **not** pre-register!



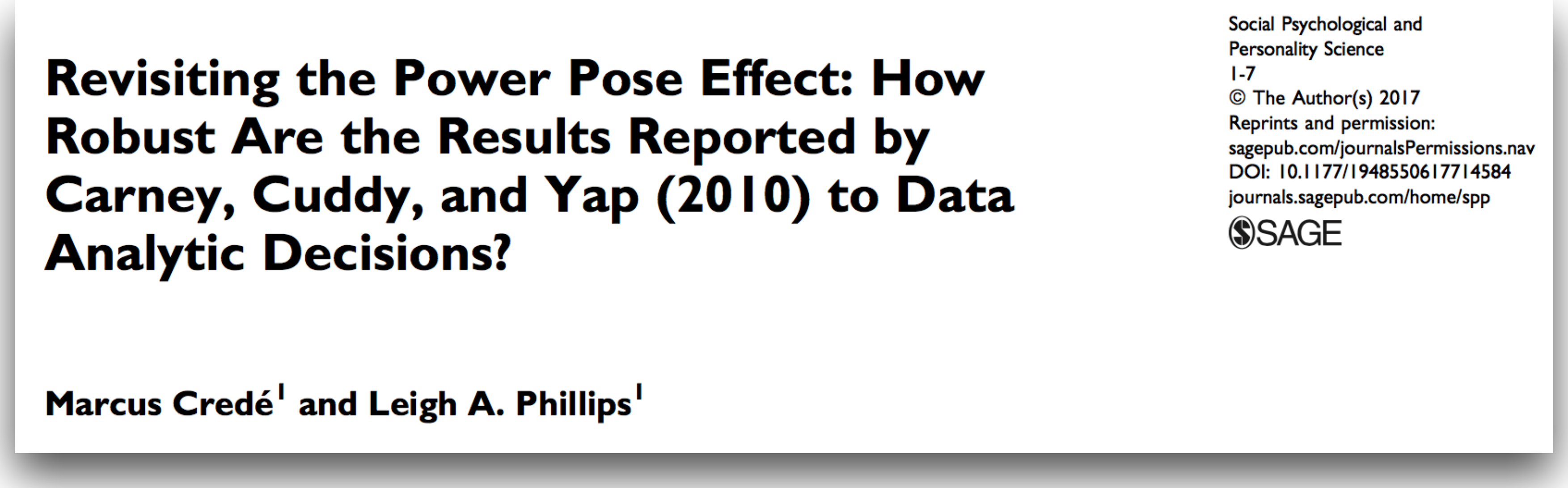
no prereg:
57%
success rate!

prereg:
8%
success rate...

http://chrisblattman.com/2016/03/01/13719/

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLoS ONE*, *10*(8), e0132382–12. http://doi.org/10.1371/journal.pone.0132382

Tool 10: Do **not** share open data



Revisiting the Power Pose Effect: How Robust Are the Results Reported by Carney, Cuddy, and Yap (2010) to Data Analytic Decisions?

Social Psychological and
Personality Science
1-7
© The Author(s) 2017
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1948550617714584
journals.sagepub.com/home/spp
SAGE

Marcus Credé[1] and Leigh A. Phillips[1]

- A "multiverse analysis" (Steegen, Tuerlinchx, Gelman, & Vanpaemel, 2016): Report results for all plausible analytical decisions
- Check robustness of results: Do several analytical paths lead to comparable conclusions?
- Based on open data by Carney et al. (2010)

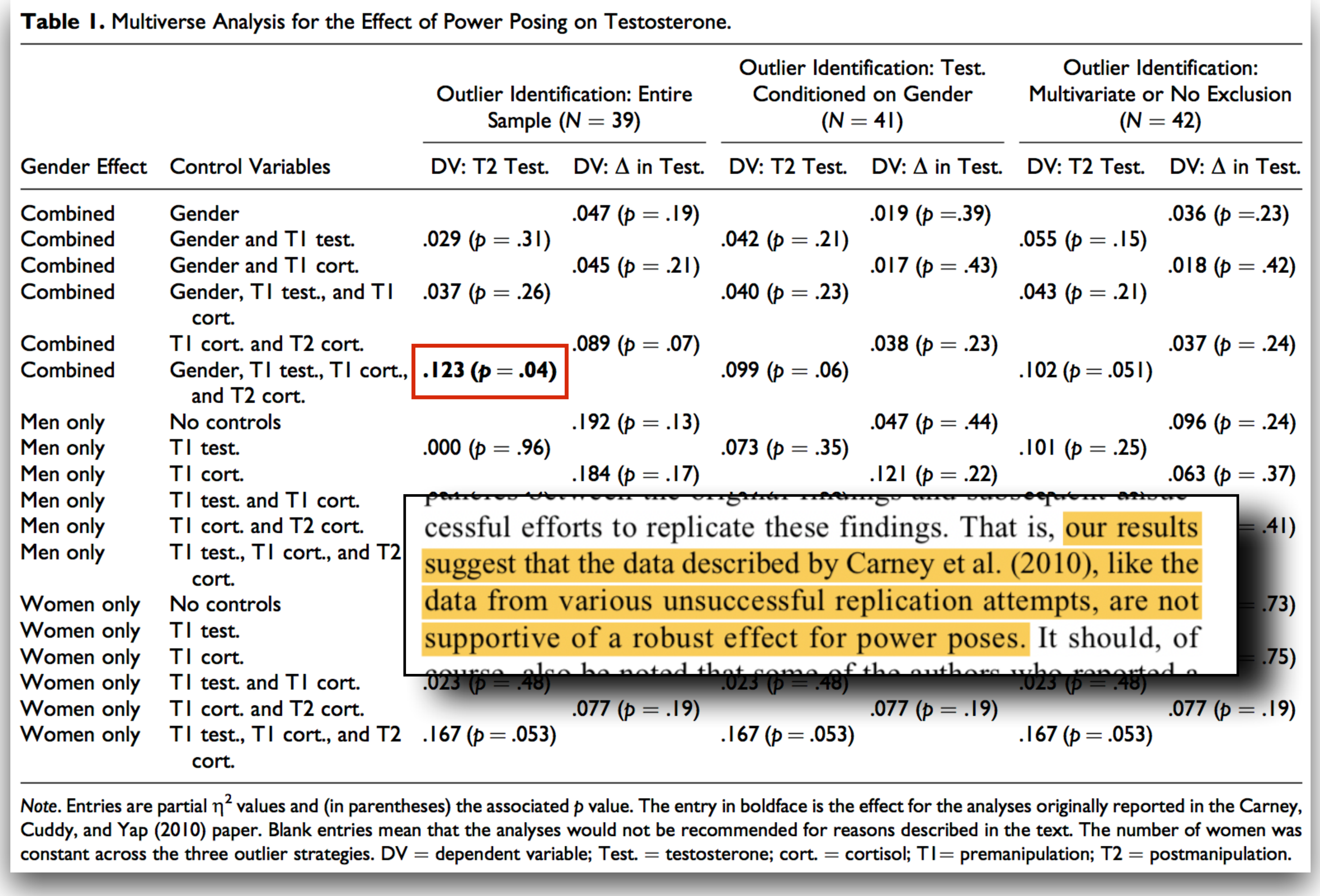


Table 1. Multiverse Analysis for the Effect of Power Posing on Testosterone.

		Outlier Identification: Entire Sample ($N = 39$)		Outlier Identification: Test. Conditioned on Gender ($N = 41$)		Outlier Identification: Multivariate or No Exclusion ($N = 42$)	
Gender Effect	Control Variables	DV: T2 Test.	DV: Δ in Test.	DV: T2 Test.	DV: Δ in Test.	DV: T2 Test.	DV: Δ in Test.
Combined	Gender		.047 ($p = .19$)		.019 ($p =.39$)		.036 ($p =.23$)
Combined	Gender and T1 test.	.029 ($p = .31$)		.042 ($p = .21$)		.055 ($p = .15$)	
Combined	Gender and T1 cort.		.045 ($p = .21$)		.017 ($p = .43$)		.018 ($p = .42$)
Combined	Gender, T1 test., and T1 cort.	.037 ($p = .26$)		.040 ($p = .23$)		.043 ($p = .21$)	
Combined	T1 cort. and T2 cort.		.089 ($p = .07$)		.038 ($p = .23$)		.037 ($p = .24$)
Combined	Gender, T1 test., T1 cort., and T2 cort.	**.123 ($p = .04$)**		.099 ($p = .06$)		.102 ($p = .051$)	
Men only	No controls		.192 ($p = .13$)		.047 ($p = .44$)		.096 ($p = .24$)
Men only	T1 test.	.000 ($p = .96$)		.073 ($p = .35$)		.101 ($p = .25$)	
Men only	T1 cort.		.184 ($p = .17$)		.121 ($p = .22$)		.063 ($p = .37$)
Men only	T1 test. and T1 cort.						
Men only	T1 cort. and T2 cort.						.41)
Men only	T1 test., T1 cort., and T2 cort.						
Women only	No controls						.73)
Women only	T1 test.						
Women only	T1 cort.						.75)
Women only	T1 test. and T1 cort.	.023 ($p = .48$)		.023 ($p = .48$)		.023 ($p = .48$)	
Women only	T1 cort. and T2 cort.		.077 ($p = .19$)		.077 ($p = .19$)		.077 ($p = .19$)
Women only	T1 test., T1 cort., and T2 cort.	.167 ($p = .053$)		.167 ($p = .053$)		.167 ($p = .053$)	

Note. Entries are partial η^2 values and (in parentheses) the associated p value. The entry in boldface is the effect for the analyses originally reported in the Carney, Cuddy, and Yap (2010) paper. Blank entries mean that the analyses would not be recommended for reasons described in the text. The number of women was constant across the three outlier strategies. DV = dependent variable; Test. = testosterone; cort. = cortisol; T1= premanipulation; T2 = postmanipulation.

parities between the original findings and subsequent unsuccessful efforts to replicate these findings. That is, our results suggest that the data described by Carney et al. (2010), like the data from various unsuccessful replication attempts, are not supportive of a robust effect for power poses. It should, of course, also be noted that some of the authors who reported a

Of 54 plausible analyses exactly ***one*** was significant.
Guess which has been reported in the original paper?

* *My lawyer told me to show that.*

Disclaimer*

- *p*-hacking increases the false positive rate
- *p*-hacking „renders the reported *p*-values essentially uninterpretable“ (ASA statement)
- *p*-hacking is ethically wrong and violates rules of good scientific practice
- If you *p*-hack systematically:
  - many of your research results will simply be wrong (depending on the prior probability of your hypotheses)
  - consequentially, your research won’t replicate
- Every time you *p*-hack, you waste public money, you waste participants’ time, you bias the literature, and **a kitten dies****.

** *If your research is about feline drug development*

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. American Statistician, 00–00. http://doi.org/10.1080/00031305.2016.1154108