12th July 2018, EADM Summer School, Salzburg

Open science practices for future-proof research

Anne Scheel

a.m.scheel@tue.nl

@AnneMScheel

I. The Replication Crisis

II. Open Data

III. Preregistration

I. The Replication Crisis

Science

Study Shows Some Evidence Of Human Precognitive Powers

In particular, "participants correctly identified the future position of erotic pictures"

By Clay Dillow posted Nov 8th, 2010 at 3:00pm

By **CBSNEWS** / **CBS** / *January 18, 2011, 8:58 AM*

Study: ESP May Be Real

Comment / Share / Tweet / Stumble / Email

Are humans psychic? Startling new study 'proves' that we can see the future

By FIONA MACRAE FOR THE DAILY MAIL
UPDATED: 09:20 BST, 18 November 2010

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407-425.

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect

Stuart J. Ritchie, Richard Wiseman, Christopher C. French

Published: March 14, 2012 • https://doi.org/10.1371/journal.pone.0033423

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

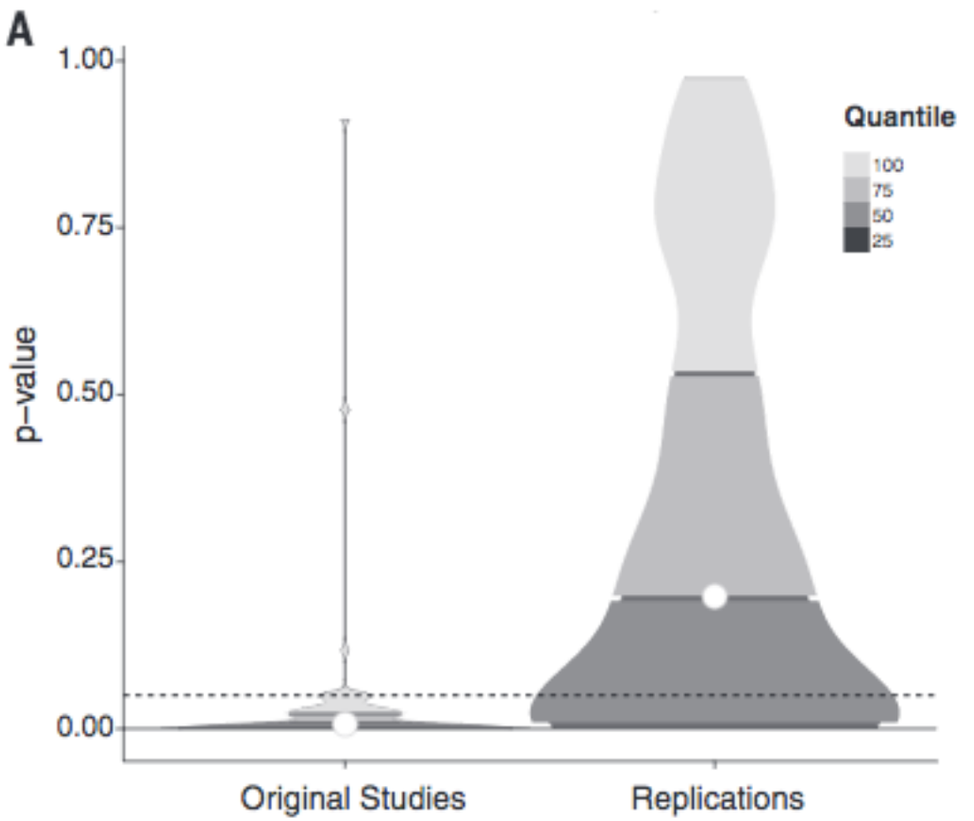
Open Science Collaboration*

100 studies from 2008 replicated

Psychological Science

Journal of Personality and Social Psychology

Journal of Experimental Psychology: Learning, Memory, and Cognition



36% replicated

Large-Scale Projects

Registered Replication Reports (RRRs)

Many Labs

Many Babies

Psychological Science Accelerator (PSA)

ManyLabs 1

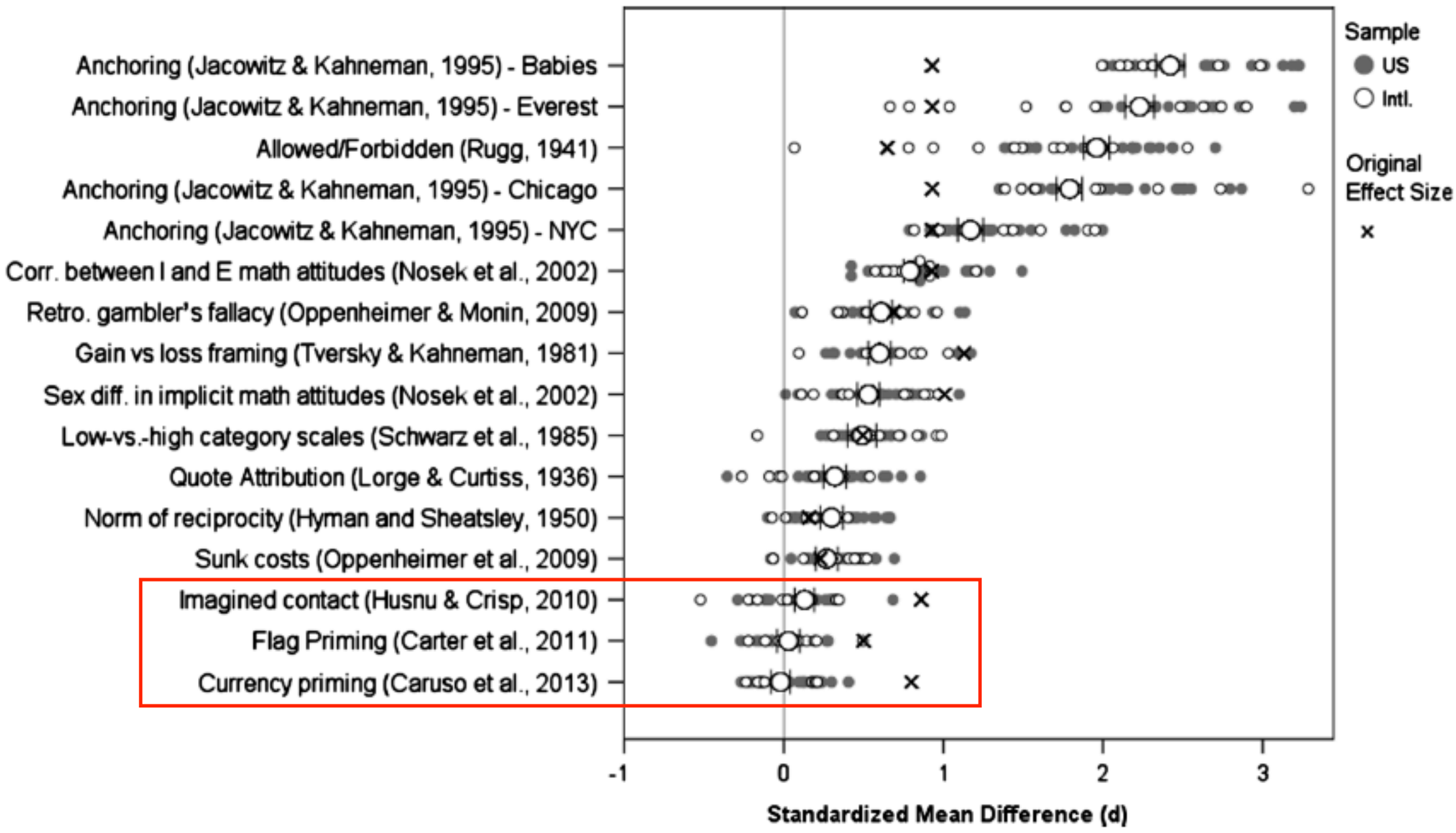
Replication

Investigating Variation in Replicability

A "Many Labs" Replication Project

Richard A. Klein,[1] Kate A. Ratliff,[1] Michelangelo Vianello,[2] Reginald B. Adams Jr.,[3]
Štěpán Bahník,[4] Michael J. Bernstein,[5] Konrad Bocian,[6] Mark J. Brandt,[7] Beach Brooks,[1]
Claudia Chloe Brumbaugh,[8] Zeynep Cemalcilar,[9] Jesse Chandler,[10,36] Winnee Cheong,[11]
William E. Davis,[12] Thierry Devos,[13] Matthew Eisner,[10] Natalia Frankowska,[6] David Furrow,[15]
Elisa Maria Galliani,[2] Fred Hasselman,[16,37] Joshua A. Hicks,[12] James F. Hovermale,[17]
S. Jane Hunt,[18] Jeffrey R. Huntsinger,[19] Hans IJzerman,[7] Melissa-Sue John,[20]
Jennifer A. Joy-Gaba,[17] Heather Barry Kappes,[21] Lacy E. Krueger,[18] Jaime Kurtz,[22]
Carmel A. Levitan,[23] Robyn K. Mallett,[19] Wendy L. Morris,[24] Anthony J. Nelson,[3]
Jason A. Nier,[25] Grant Packard,[26] Ronaldo Pilati,[27] Abraham M. Rutchick,[28]
Kathleen Schmidt,[29] Jeanine L. Skorinko,[20] Robert Smith,[14] Troy G. Steiner,[3] Justin Storbeck,[8]
Lyn M. Van Swol,[30] Donna Thompson,[15] A. E. van 't Veer,[7] Leigh Ann Vaughn,[31]
Marek Vranka,[32] Aaron L. Wichman,[33] Julie A. Woodzicka,[34] and Brian A. Nosek[29,35]

ManyLabs 1

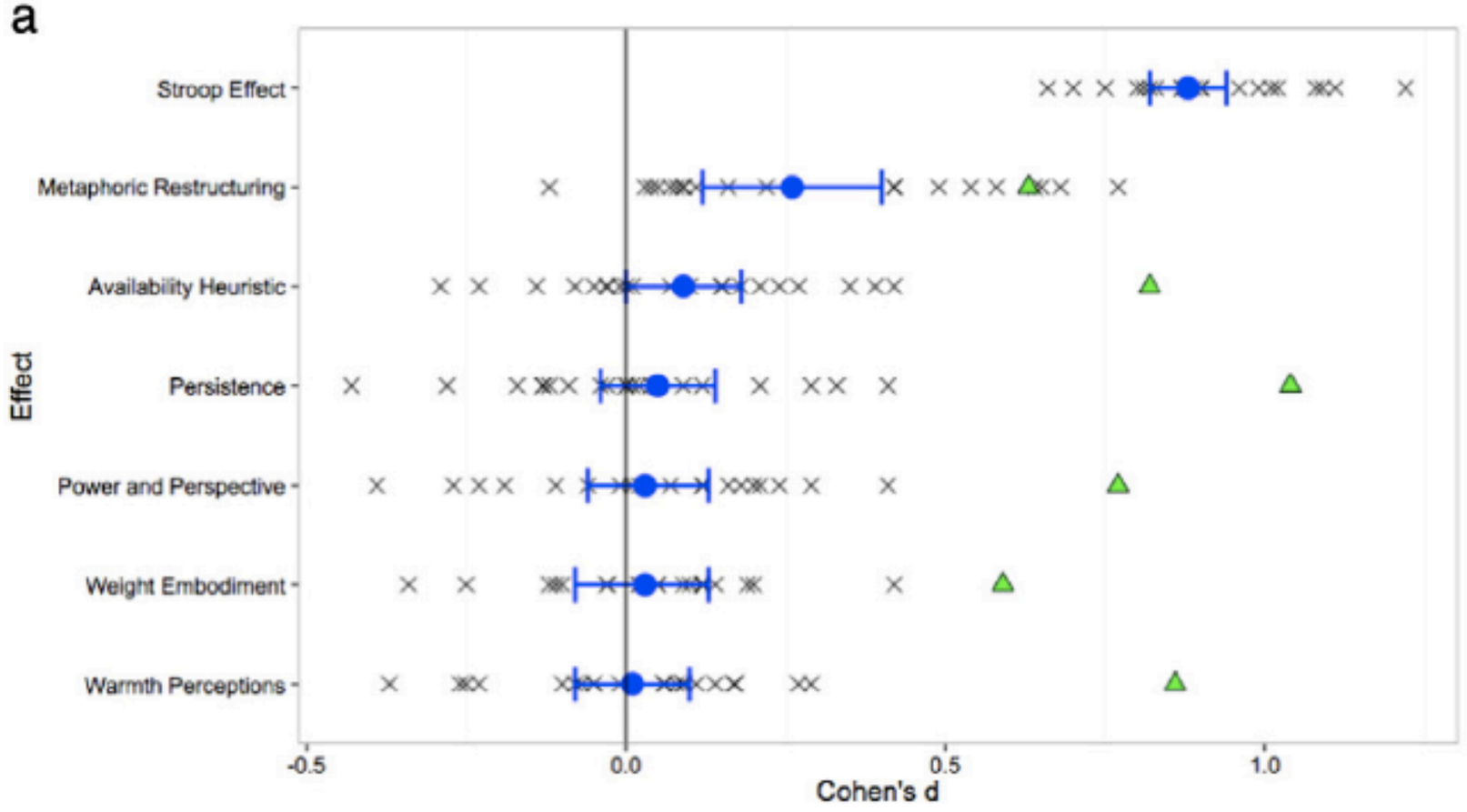

Anchoring (Jacowitz & Kahneman, 1995) - Babies
Anchoring (Jacowitz & Kahneman, 1995) - Everest
Allowed/Forbidden (Rugg, 1941)
Anchoring (Jacowitz & Kahneman, 1995) - Chicago
Anchoring (Jacowitz & Kahneman, 1995) - NYC
Corr. between I and E math attitudes (Nosek et al., 2002)
Retro. gambler's fallacy (Oppenheimer & Monin, 2009)
Gain vs loss framing (Tversky & Kahneman, 1981)
Sex diff. in implicit math attitudes (Nosek et al., 2002)
Low-vs.-high category scales (Schwarz et al., 1985)
Quote Attribution (Lorge & Curtiss, 1936)
Norm of reciprocity (Hyman and Sheatsley, 1950)
Sunk costs (Oppenheimer et al., 2009)
Imagined contact (Husnu & Crisp, 2010)
Flag Priming (Carter et al., 2011)
Currency priming (Caruso et al., 2013)
-1
0
1
2
3
Standardized Mean Difference (d)
Sample
US
Intl.
Original
Effect Size

ManyLabs 3

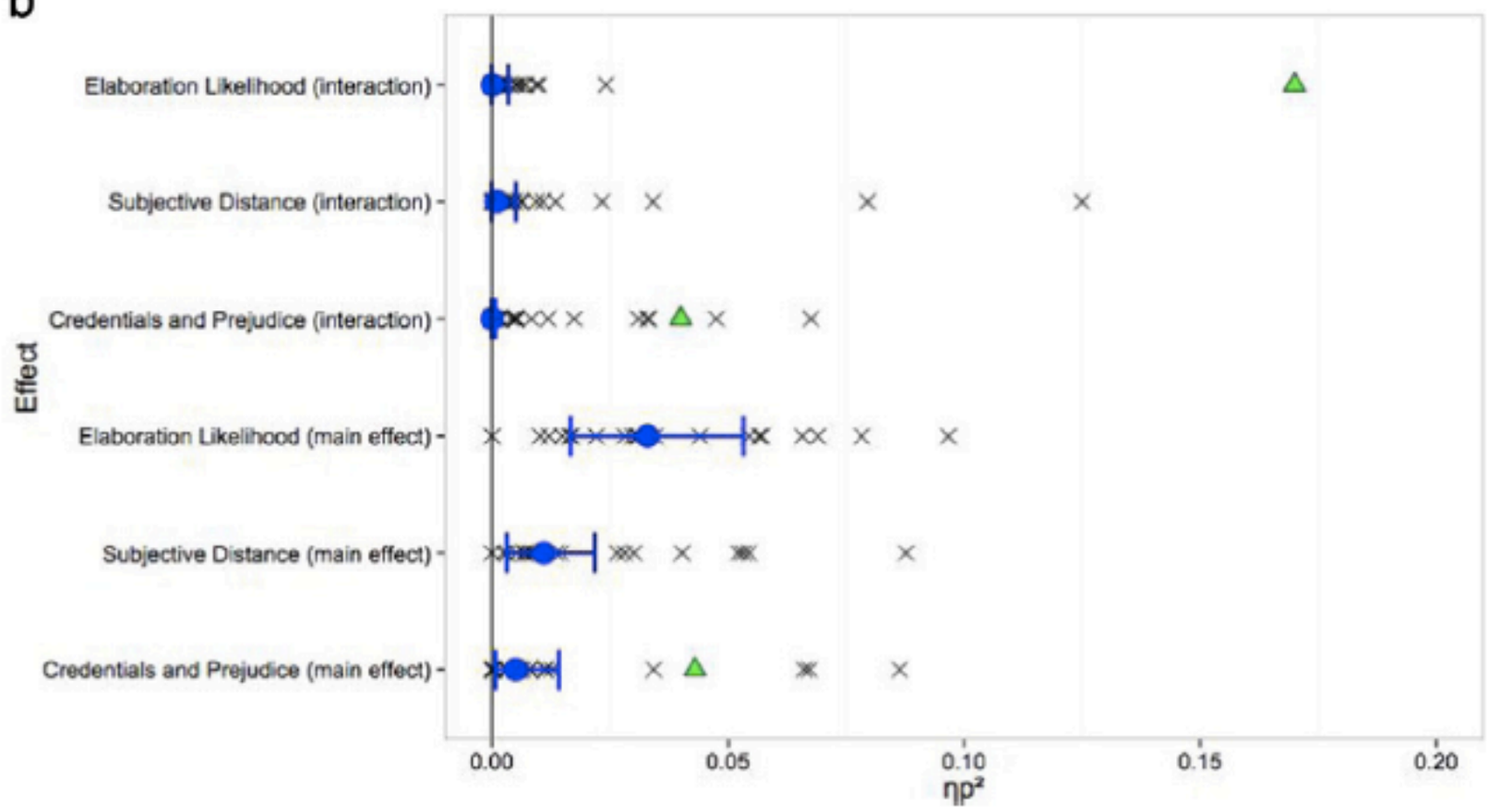
Many Labs 3: Evaluating participant pool quality across the academic semester via replication☆

Charles R. Ebersole [a,*], Olivia E. Atherton [b], Aimee L. Belanger [c], Hayley M. Skulborstad [c], Jill M. Allen [d], Jonathan B. Banks [e], Erica Baranski [f], Michael J. Bernstein [g], Diane B.V. Bonfiglio [h], Leanne Boucher [e], Elizabeth R. Brown [i], Nancy I. Budiman [b], Athena H. Cairo [j], Colin A. Capaldi [k], Christopher R. Chartier [h], Joanne M. Chung [b], David C. Cicero [l], Jennifer A. Coleman [j], John G. Conway [m], William E. Davis [n], Thierry Devos [o], Melody M. Fletcher [p], Komi German [b], Jon E. Grahe [q], Anthony D. Hermann [r], Joshua A. Hicks [n], Nathan Honeycutt [o], Brandon Humphrey [c], Matthew Janus [a], David J. Johnson [s], Jennifer A. Joy-Gaba [j], Hannah Juzeler [q], Ashley Keres [h], Diana Kinney [a], Jacqeline Kirshenbaum [r], Richard A. Klein [m], Richard E. Lucas [s], Christopher J.N. Lustgraaf [p], Daniel Martin [a], Madhavi Menon [e], Mitchell Metzger [h], Jaclyn M. Moloney [j], Patrick J. Morse [f], Radmila Prislin [o], Timothy Razza [e], Daniel E. Re [t], Nicholas O. Rule [t], Donald F. Sacco [p], Kyle Sauerberger [f], Emily Shrider [h], Megan Shultz [q], Courtney Siemsen [r], Karin Sobocko [k], R. Weylin Sternglanz [e], Amy Summerville [c], Konstantin O. Tskhay [t], Zack van Allen [k], Leigh Ann Vaughn [u], Ryan J. Walker [c], Ashley Weinberg [o], John Paul Wilson [t], James H. Wirth [v], Jessica Wortman [s], Brian A. Nosek [w]

a


Stroop Effect
Metaphoric Restructuring
Availability Heuristic
Persistence
Power and Perspective
Weight Embodiment
Warmth Perceptions
Effect
-0.5
0.0
0.5
1.0
Cohen's d

b

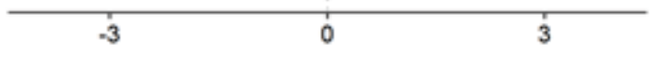

Elaboration Likelihood (interaction)
Subjective Distance (interaction)
Credentials and Prejudice (interaction)
Elaboration Likelihood (main effect)
Subjective Distance (main effect)
Credentials and Prejudice (main effect)
Effect
0.00
0.05
0.10
0.15
0.20
ηp²

Registered Replication Reports (RRRs)

effect	original	replication	*N*	result
verbal overshadowing	Schooler & Engstler-Schooler (1990)	Alogna et al. (2014)	2566 / 1491	(✓)
verb aspect & intent attribution	Hart & Albarracín (2011)	Eerland et al. (2016)	1366	–
ego depletion	Sripada et al. (2014)	Hagger et al. (2016)	2141	–
relationship commitment & forgiveness	Finkel et al. (2014)	Cheung et al. (2016)	2284	–/?
facial feedback	Strack et al. (1988)	Wagenmakers et al. (2016)	1894	–
cooperation & time pressure	Rand et al. (2012)	Bouwmeester et al. (2017)	2163	–

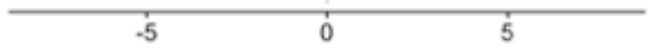
RRR 9: Hostility priming (Srull & Wyer, 1979)

Lab	20% Hostile Prime	N	80% Hostile Prime	N	Mean Difference (80% Hostile Prime - 20% Hostile Prime)	ES	95% CI
Srull & Wyer (1979)	5.56	8	8.57	8		3.01	
BIRT	7.06	102	7.61	103		0.55	[0.15, 0.94]
ACZEL	6.19	116	6.50	109		0.32	[-0.03, 0.67]
SUCHOTZKI	7.37	125	7.65	121		0.29	[-0.01, 0.58]
FERREIRA-SANTOS	6.97	130	7.25	104		0.28	[-0.09, 0.66]
WICK	6.90	170	7.16	173		0.26	[-0.05, 0.58]
HOLZMEISTER	7.21	128	7.42	125		0.22	[-0.16, 0.59]
LOSCHELDER	7.26	112	7.43	114		0.17	[-0.17, 0.51]
MCCARTHY	7.35	125	7.51	154		0.16	[-0.19, 0.51]
KLEIN SELLE & ROZMANN	7.13	142	7.28	157		0.14	[-0.16, 0.44]
KOPPEL	7.20	120	7.33	122		0.13	[-0.23, 0.48]
ACAR	6.66	108	6.72	106		0.07	[-0.37, 0.50]
MEIJER	6.92	170	6.99	178		0.07	[-0.20, 0.33]
OZDOGRU	6.58	171	6.61	161		0.03	[-0.30, 0.36]
ROETS	7.06	100	7.08	104		0.02	[-0.30, 0.34]
GONZALEZ-IRAIZOZ	7.31	116	7.33	113		0.02	[-0.31, 0.34]
VANPAEMEL	7.03	116	7.04	121		0.01	[-0.28, 0.31]
WIGGINS	6.96	124	6.95	120		-0.01	[-0.41, 0.39]
LAINE	6.19	125	6.18	128		-0.01	[-0.34, 0.31]
SUTAN	6.02	129	5.99	123		-0.03	[-0.36, 0.30]
EVANS	7.29	124	7.18	119		-0.11	[-0.49, 0.27]
VERSCHUERE	7.14	146	6.89	139		-0.26	[-0.54, 0.03]
PENNINGTON	7.60	114	7.31	103		-0.29	[-0.63, 0.05]
Meta-analytic average for Trait Ratings	6.97	2813	7.06	2797		0.08	[0.00, 0.16]

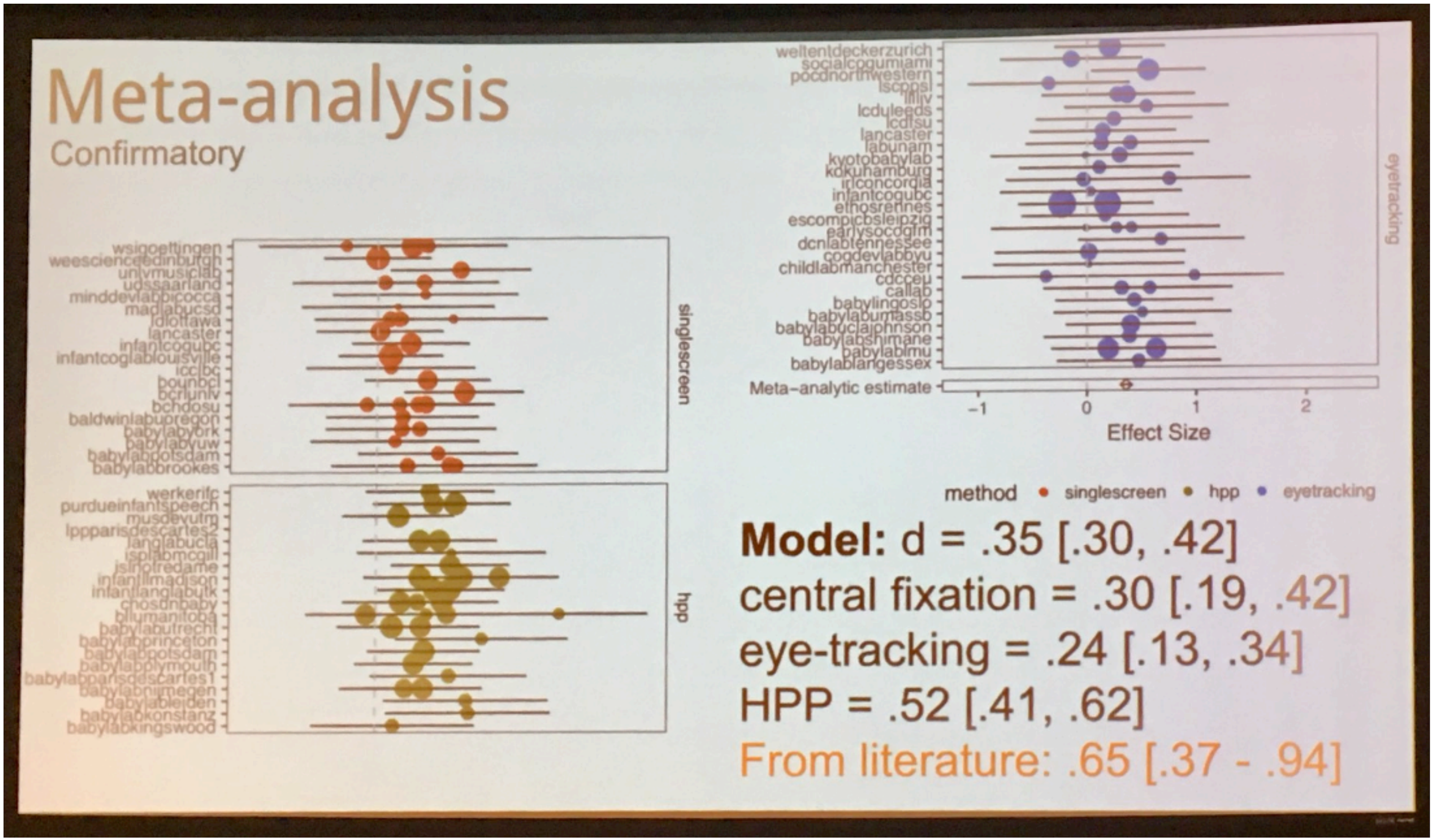


RRR10: 10 commandments priming (Mazar, Amir, & Ariely, 2008)

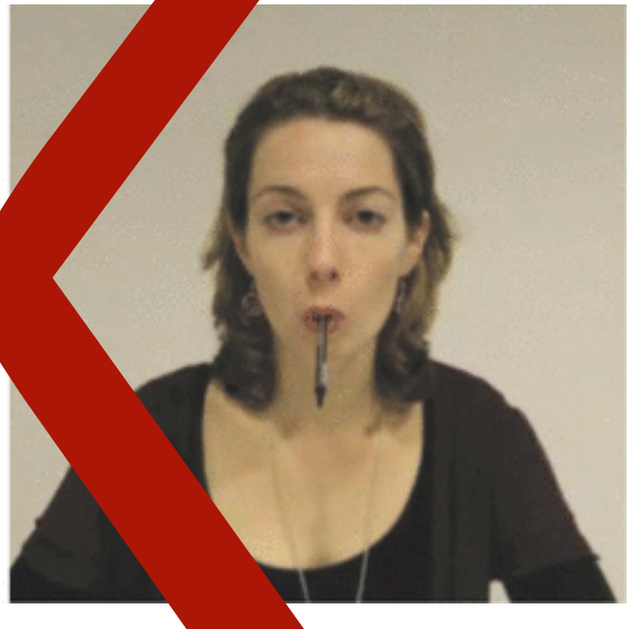
Lab	Commandments	N	Books	N	Matrices Solved (Cheat)	ES	95% CI
Original result	2.77	56	4.22	50		-1.45	[-2.61, -0.29]
Laine	2.19	53	2.56	61		-0.37	[-1.18, 0.44]
klein Selle & Rozmann	3.31	70	3.58	69		-0.27	[-1.11, 0.58]
Aczel	3.18	55	3.44	54		-0.26	[-1.22, 0.69]
Ferreira-Santos	2.67	51	2.85	54		-0.19	[-1.14, 0.77]
Meijer	3.02	84	3.18	84		-0.15	[-0.75, 0.44]
Loschelder	3.98	59	4.09	55		-0.11	[-0.86, 0.65]
Wick	3.19	81	3.28	79		-0.09	[-1.06, 0.87]
Suchotzki	3.83	60	3.83	60		0.00	[-0.93, 0.93]
Sutan	4.68	34	4.66	87		0.02	[-0.79, 0.83]
Vanpaemel	3.61	56	3.44	59		0.17	[-0.55, 0.88]
Verschuere	3.73	64	3.55	69		0.18	[-0.55, 0.91]
Wiggins	2.34	62	2.15	61		0.19	[-0.51, 0.90]
González-Iraizoz	3.67	55	3.41	58		0.26	[-0.77, 1.28]
Koppel	3.11	62	2.73	62		0.39	[-0.40, 1.17]
Birt	3.04	53	2.63	49		0.41	[-0.58, 1.39]
McCarthy	3.40	50	2.83	46		0.57	[-0.87, 2.02]
Evans	3.08	53	2.23	57		0.85	[-0.13, 1.83]
Holzmeister	7.02	60	5.91	64		1.11	[-0.30, 2.52]
Özdoğru	4.45	38	3.26	66		1.19	[0.01, 2.37]
Meta-analytic average for 10 commandments effect	3.55	1100	3.35	1194		0.11	[-0.09, 0.31]



ManyBabies 1




LOW
ELDERL



PSY 607: Everything is Fucked

Prof. Sanjay Srivastava

Class meetings: Mondays 9:00 – 10:50 in 257 Straub

Office hours: Held on Twitter at your convenience (@hardsci)

Roots of the Replication Crisis

Publication Bias

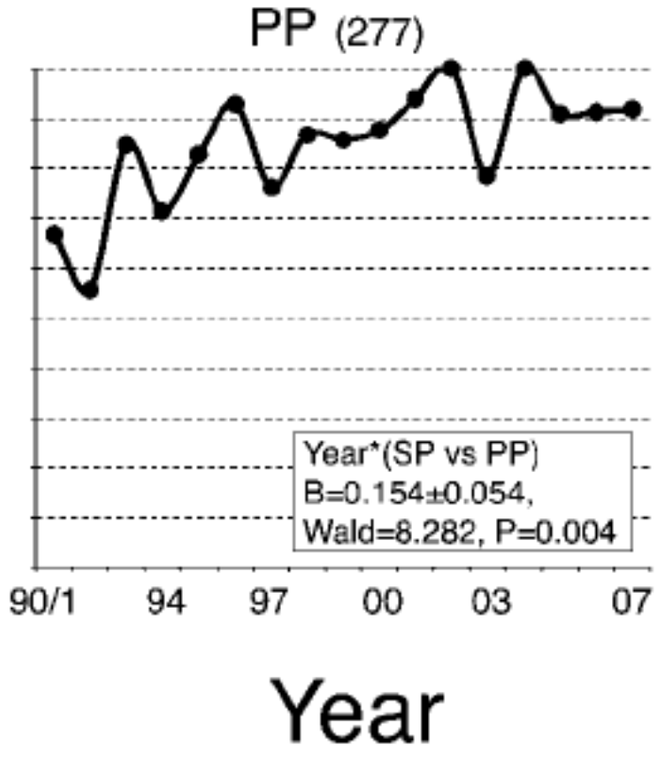
TABLE 32

PER CENT OF ARTICLES USING TESTS OF SIGNIFICANCE AND PER CENT OF ARTICLES REJECTING H_0

Journals: All Issues From January to December	Per Cent of Articles Using Tests of All Articles Published (2/1)	Per Cent of Articles Rejecting H_0 of All Articles Using Tests (3/2)	Per Cent of Articles Not Rejecting H_0 of All Articles Using Tests (4/2)
Experimental Psychology (1955)	85.48	99.06	0.94
Comparative and Physiological Psychology (1956)	79.66	96.81	3.19
Clinical Psychology (1955)	76.54	95.16	4.84
Social Psychology (1955)	82.05	96.88	3.12
Total	81.22	97.28	2.72

Sterling, T. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association, 54*(285), 30-34.

Publication Bias



Psychology: ~92% significant results

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90,* 891-904.

Publication Bias: Amplified by Citation

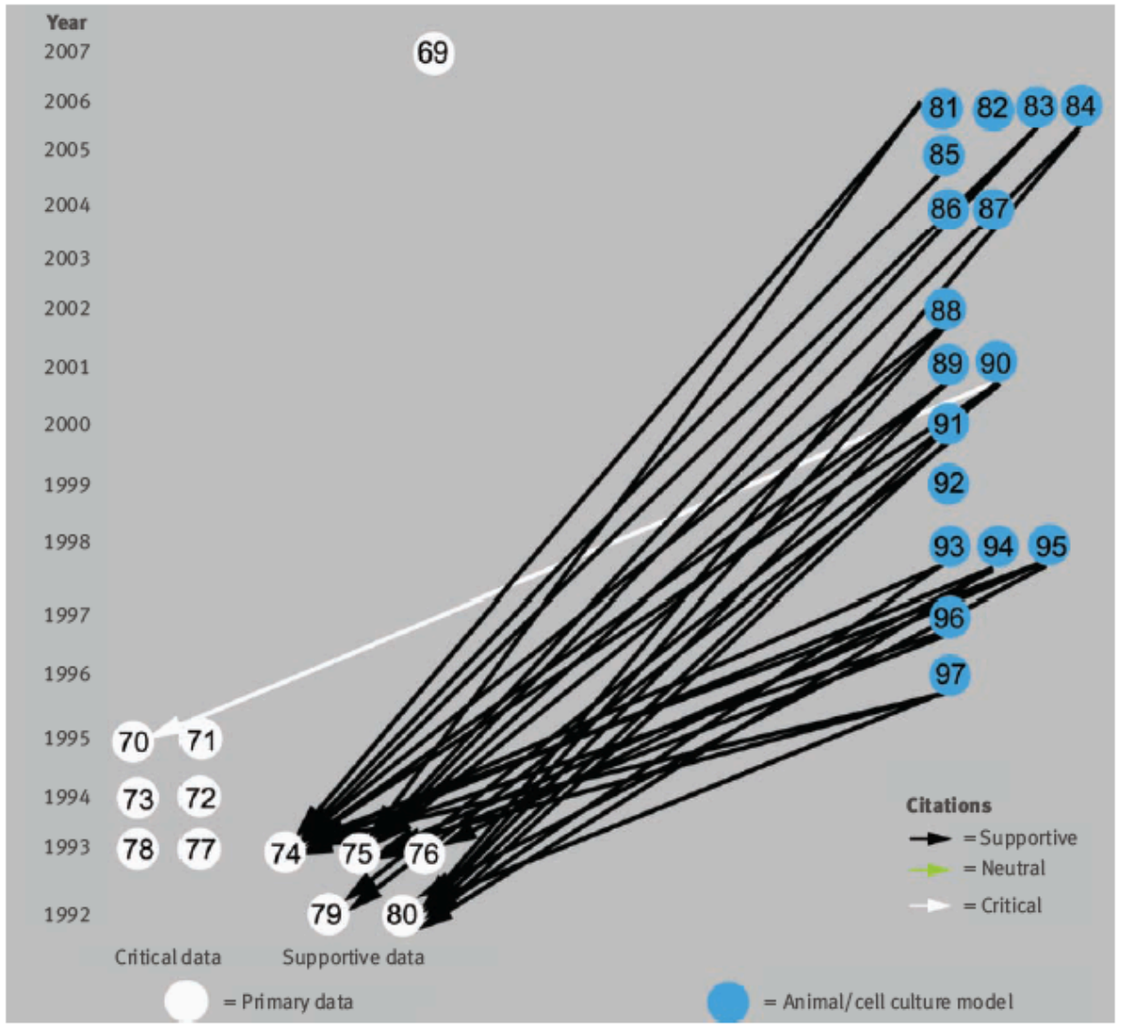
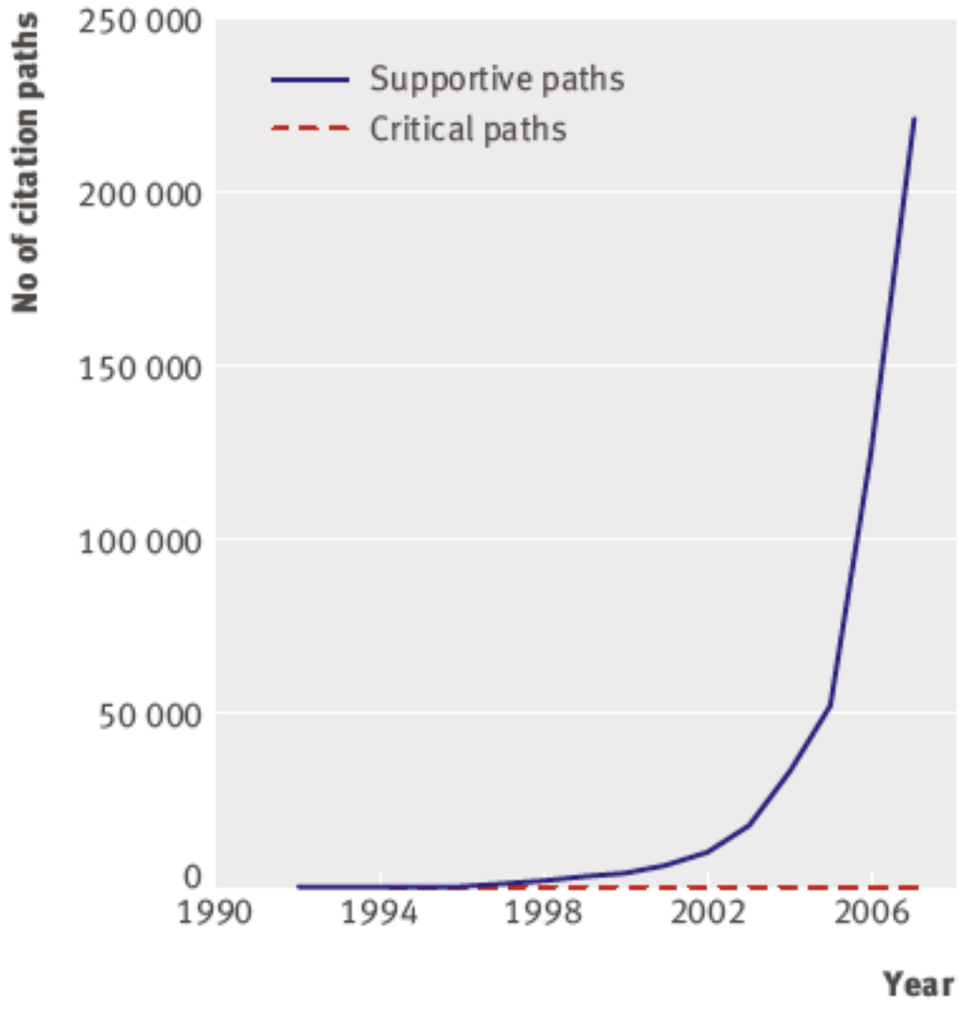
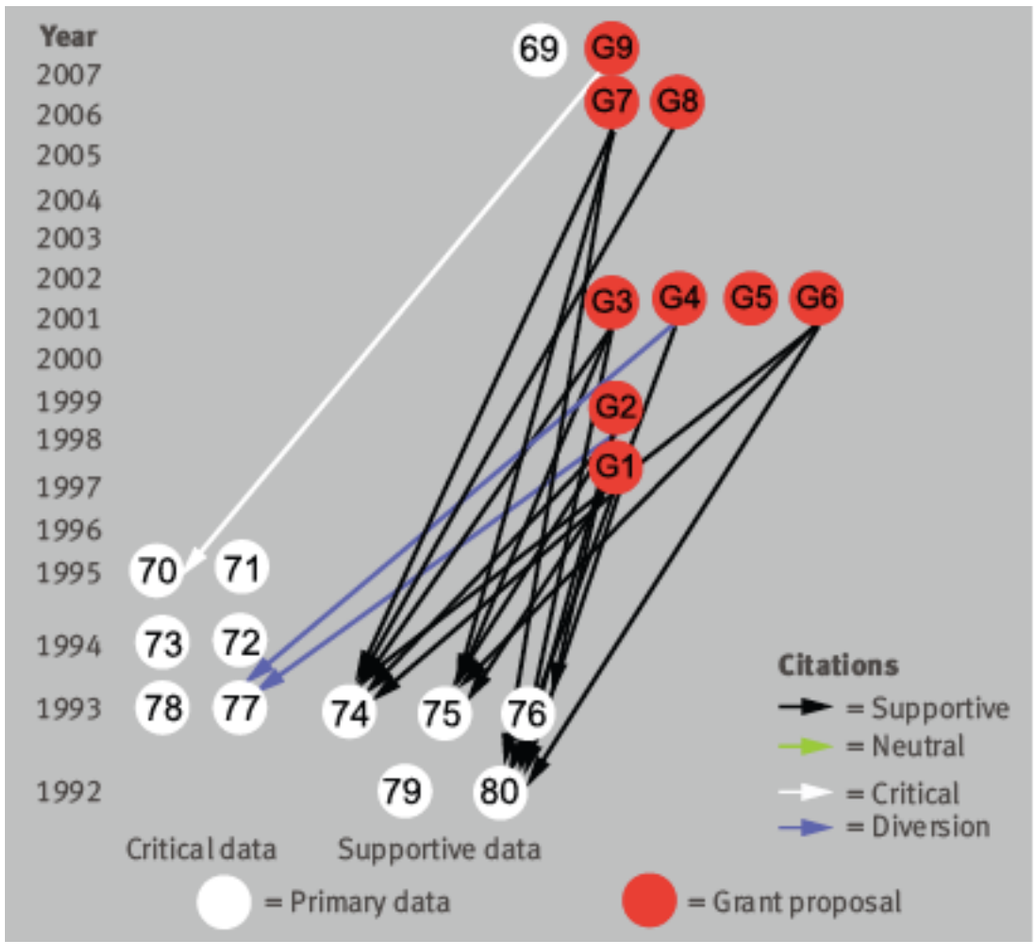


Fig 3| Citations from animal and cell culture model papers to primary data papers supporting rationale for overproduction of β amyloid precursor protein mRNA as a valid model of inclusion body myositis. Only one of 32 citations flows to papers[70-73 77 78] that present data that conflict with the validity of these models



Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *BMJ, 339*, b2680.

Publication Bias: Amplified by Citation



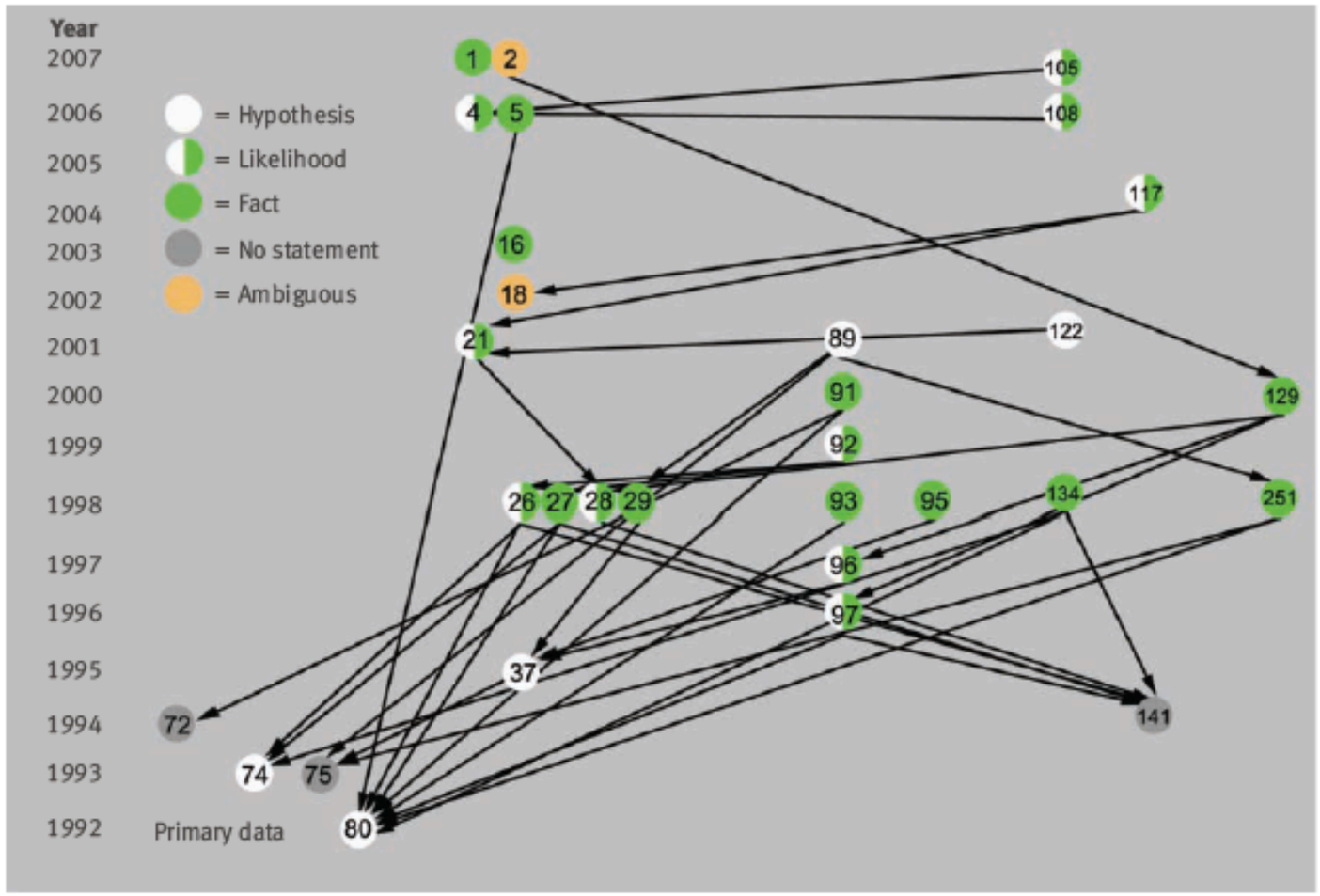


Fig 5 | Conversion of hypothesis to fact through citation alone. Citations on statement that accumulation of β amyloid "precedes" other abnormalities in inclusion body myositis muscle. Statement as fact is supported through citation to papers that only state it as hypothesis (for example, references 5 to 80, 91 to 80, 134 to 74) or sometimes supported by citation to papers that contain no statements addressing it (for example, references 91 to 72, 251 to 75; dead end citations). This phenomenon might be called citation transmutation (see web extra note 10 for statements)

Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *BMJ, 339,* b2680.

Low Power (+publication bias)

		Small effect		Medium effect		Large effect	
Subfields or other surveys	**Records/Articles**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
Cognitive neuroscience	7,888/1,192	0.11	0.14	0.40	0.44	0.70	0.67
Psychology	16,887/2,261	0.16	0.23	0.60	0.60	0.81	0.78
Medical	2,066/348	0.15	0.23	0.59	0.57	0.80	0.77
All subfields	26,841/3,801	0.11	0.17	0.44	0.49	0.73	0.71
Cohen (1962)	2,088/70	0.17	0.18	0.46	0.48	0.89	0.83
Sedlmeier & Gigerenzer (1989)	54 articles	0.14	0.21	0.44	0.50	0.90	0.84
Rossi (1990)	6,155/221	0.12	0.17	0.53	0.57	0.89	0.83
Rossi (1990); means of surveys	25 surveys		0.26		0.64		0.85

doi:10.1371/journal.pbio.2000797.t001

Szucs,, D. & Ioannidis. J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15(3),* e2000797.

Low Power (+publication bias)

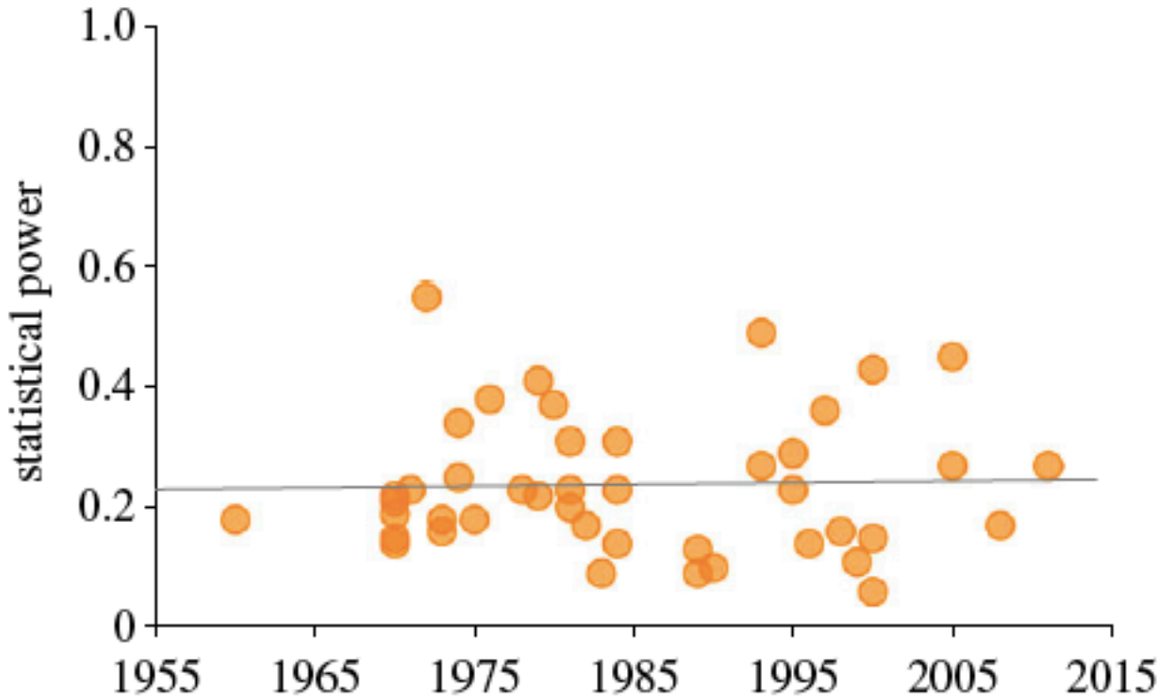


Figure 1. Average statistical power from 44 reviews of papers published in journals in the social and behavioural sciences between 1960 and 2011. Data are power to detect small effect sizes ($d = 0.2$), assuming a false-positive rate of $\alpha = 0.05$, and indicate both very low power (mean = 0.24) but also no increase over time ($R^2 = 0.00097$).

Smaldino, P. E. & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3*, 160384.

Low Power (+publication bias)

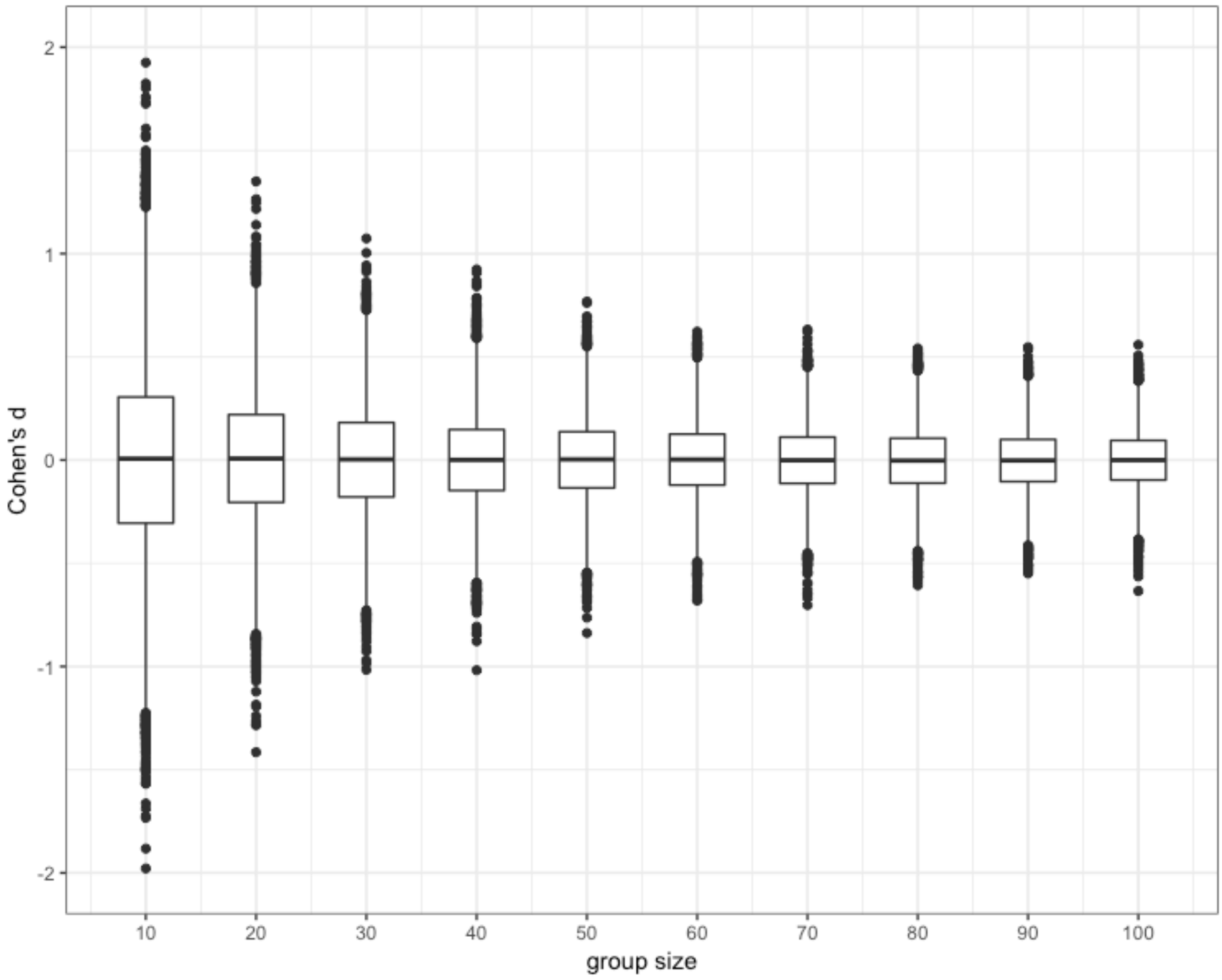
"The persistence of poor methods results partly from incentives that favour them, leading to the **natural selection of bad science**. This dynamic requires no conscious strategizing—no deliberate cheating nor loafing—by scientists, only that publication is a principal factor for career advancement. Some normative methods of analysis have almost certainly been selected to further publication instead of discovery. (...)

As in the real world, successful labs produce more 'progeny,' such that their methods are more often copied and their students are more likely to start labs of their own. **Selection for high output leads to poorer methods and increasingly high false discovery rates.**"

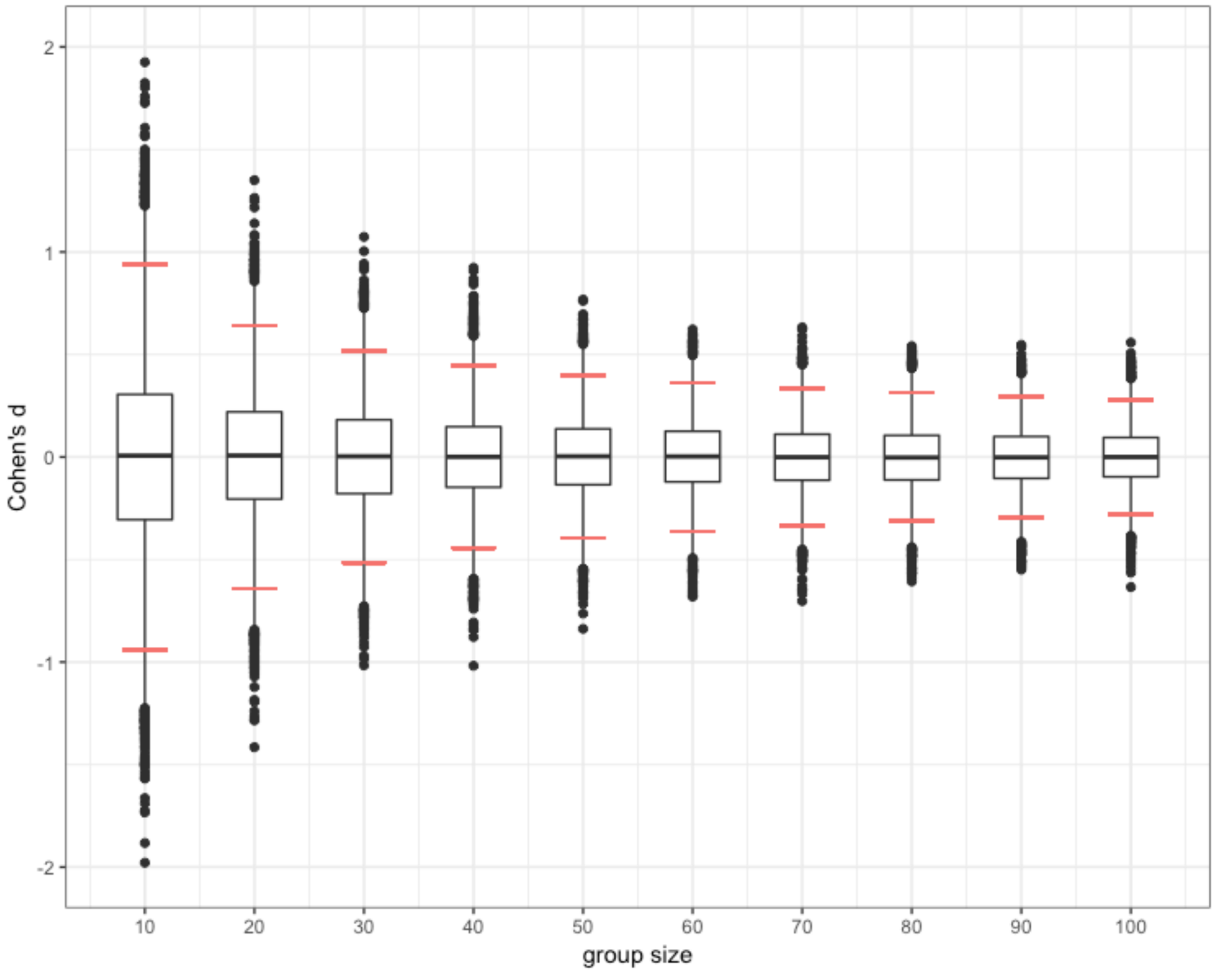
Smaldino, P. E. & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3,* 160384.

Exercise 1: publication bias

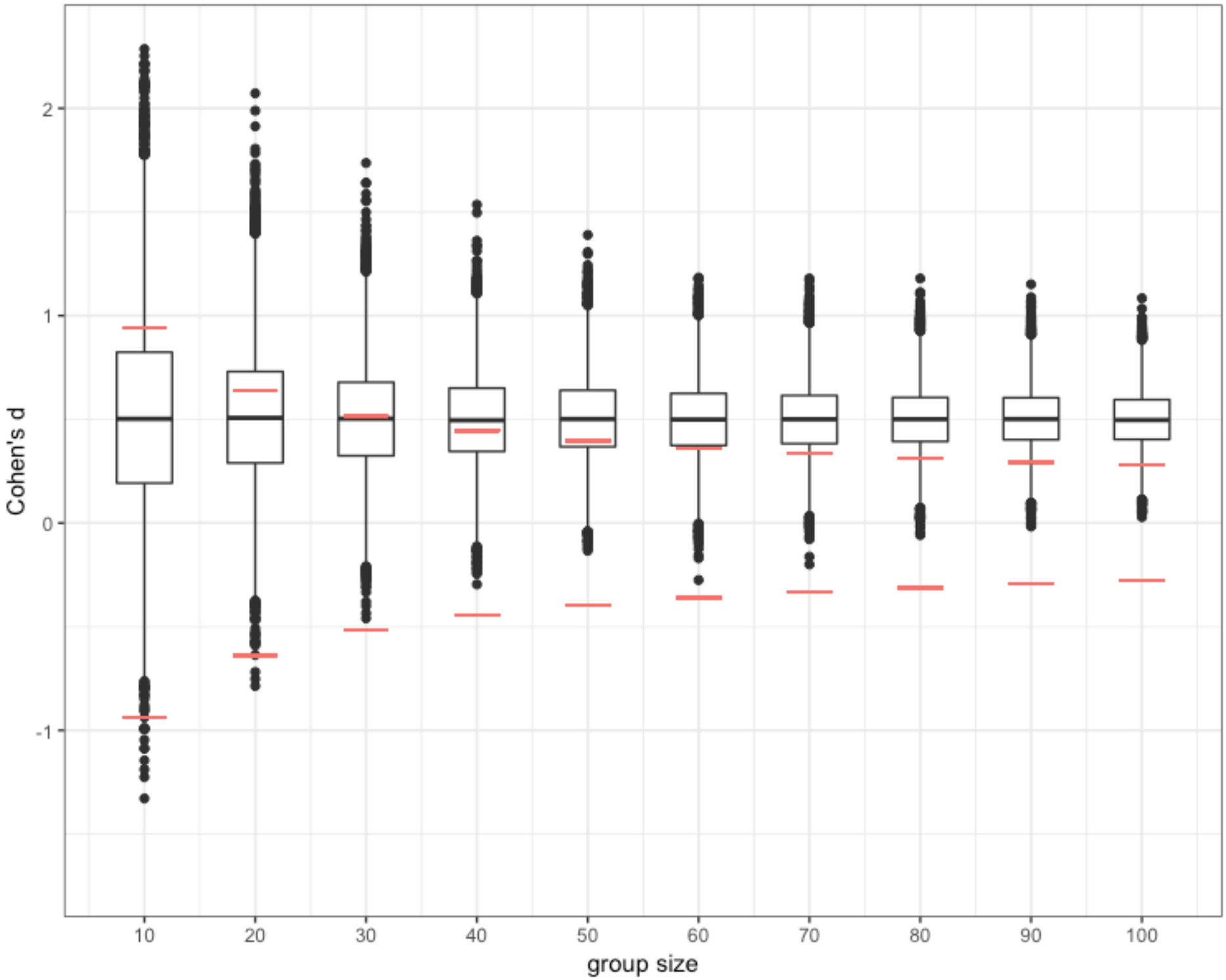
10,000 two-sample t-tests, d = 0



10,000 two-sample t-tests, d = 0



10,000 two-sample t-tests, d = 0.5



Researcher Degrees of Freedom/ Questionable Research Practices

Study 2: musical contrast and chronological rejuvenation

Using the same method as in Study 1, we asked 20 University of Pennsylvania undergraduates to listen to either "When I'm Sixty-Four" by The Beatles or "Kalimba." Then, in an ostensibly unrelated task, they indicated their birth date (mm/dd/yyyy) and their father's age. We used father's age to control for variation in baseline age across participants.

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted $M = 20.1$ years) rather than to "Kalimba" (adjusted $M = 21.5$ years), $F(1, 17) = 4.92, p = .040$.

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*(11), 1359-1366.

Researcher Degrees of Freedom

Table 3. Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

Using the same method as in Study 1, we asked ~~20~~ 34 University of Pennsylvania undergraduates to listen only **to either "When I'm Sixty-Four" by The Beatles or "Kalimba"** or "Hot Potato" by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. **Then, in an ostensibly unrelated task, they indicated** only **their birth date (mm/dd/yyyy) and** how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with "computers are complicated machines," **their father's age,** their mother's age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as "the good old days," and their gender. **We used father's age to control for variation in baseline age across participants.**

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted *M* = 20.1 years) rather than to "Kalimba" (adjusted *M* = 21.5 years), *F*(1, 17) = 4.92, *p* = .040. Without controlling for father's age, the age difference was smaller and did not reach significance (*M*s = 20.3 and 21.2, respectively), *F*(1, 18) = 1.01, *p* = .33.

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*(11), 1359-1366.

Researcher Degrees of Freedom

"*p*-hacking"

- stop data collection early or add participants depending on the data
- change inclusion/exclusion criteria depending on the data
- drop conditions/variables that "didn‘t work"
- selectively report/focus on outcome variables that "worked"
- add or drop covariates depending on the data
- use a one-sided test although sampling was planned for a two-sided test

Exercise 2: *p*-curve

Power Posing

second most-watched TED
talk of all time
(44+ million views)

www.ted.com/talks/
amy_cuddy_your_body_lang
uage_shapes_who_you_are

729 citations since 2010

	Carney, Cuddy, & Yap (2010)	**Ranehill et al. (2015)**	**CRSP special issue (2017)**
	N = 42	N = 200	7 studies, total N > 1000
Increased feelings of power (self-reported)	✔	✔	✔**... maybe?**
Increased testosterone levels	×	×	×
Decreased cortisol levels	×	×	×
Increased financial risk-taking	×	×	×

- Bailey, LaFrance, and Dovidio (2017) sought to investigate an interaction of power posing, target gender, and participant gender. They did not replicate the effect of power poses on risky behavior.
- Bombari, Schmid Mast, and Pulfrey (2017) planned to test whether imagined or performed power poses had similar effects. They did not replicate the effect of power poses on risky behavior.
- Klaschinski, Schnabel, and Schröder-Abé (2017) wanted to replicate the effects of power posing on dominance and social sensitivity in an interview context, but they did not replicate the effects.
- Jackson, Nault, Smart Richman, LaBelle, and Rohleder (2017) sought to test the effect of power posing on self-concept. Although a preliminary study obtained an interesting effect, they did not replicate this in the higher-powered, preregistered study.
- Keller, Johnson, and Harder (2017) wanted to test whether awareness of the function of power poses moderates their effectiveness. They did not replicate the basic power pose effect.
- Latu, Duffy, Pardal, and Alger (2017) tested an interesting dependent variable in the context of power poses, persuasive messages. They did not observe any effect of power poses on persuasive message perception.
- Ronay, Tybur, van Huijstee, and Morssinkhoff (2017) wanted to investigate the mediating role of testosterone and overconfidence on the link between power posing and risk taking, but they did not replicate the effect.

Carney: My position on "Power Poses"

http://faculty.haas.berkeley.edu/dana_carney/pdf_My_position_on_power_poses.pdf

3. The sample size is tiny.

5. Initially, the primary DV of interest was risk-taking. We ran subjects in chunks and checked the effect along the way. It was something like 25 subjects run, then 10, then 7, then 5. Back then this did not seem like p-hacking. It seemed like saving money (assuming your effect size was big enough and p-value was the only issue).

6. Some subjects were excluded on bases such as "didn't follow directions." The total number of exclusions was 5. The final sample size was N= 42.

8. For the risk-taking DV: One p-value for a Pearson chi square was .052 and for the Likelihood ratio it was .05. The smaller of the two was reported despite the Pearson being the more ubiquitously used test of significance for a chi square. This is clearly using a "researcher degree of freedom." I had found evidence that it is more appropriate to use "Likelihood" when one has smaller samples and this was how I convinced myself it was OK.

9. For the Testosterone DV: An outlier for testosterone were found. It was a clear outlier (+ 3SDs away from the mean). Subjects with outliers were held out of the hormone analyses but not all analyses.

10. The self-report DV was p-hacked in that many different power questions were asked and those chosen were the ones that "worked."

http://faculty.haas.berkeley.edu/dana_carney/pdf_My_position_on_power_poses.pdf

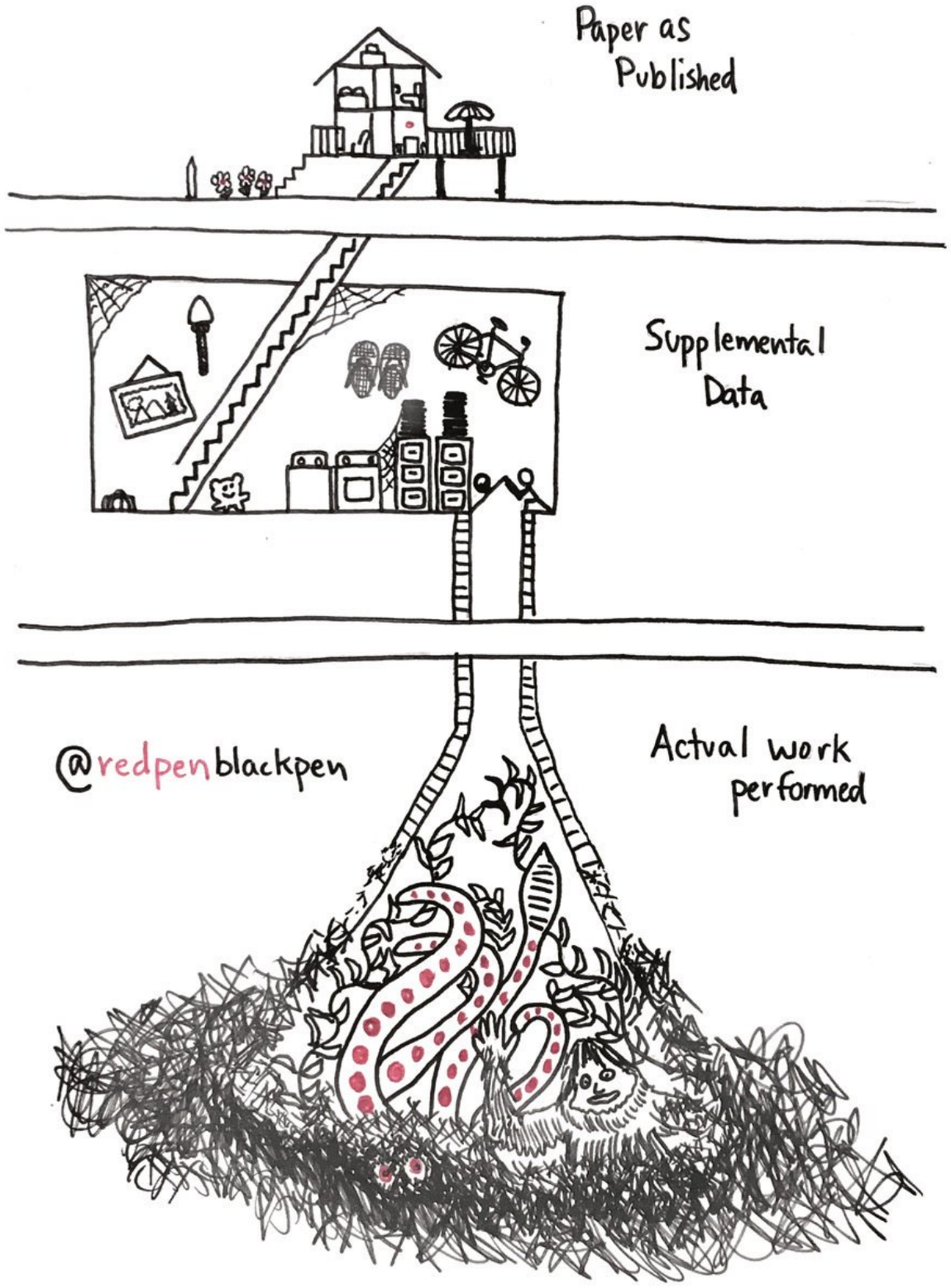
Confounds in the Original Paper (Which should have been evident in 2010 but only in hindsight did these confounds become so obviously clear):

1. The experimenters were both aware of the hypothesis. The experimenter who ran the pilot study was less aware but by the end of running the experiment certainly had a sense of the hypothesis. The experimenters who ran the main experiment (the experiment with the hormones) knew the hypothesis.

2.When the risk-taking task was administered, participants were told immediately after whether they had "won." Winning included an extra prize of $2 (in addition to the $2 they had already received). Research shows that winning increases testosterone (...)

3. Gender was not dealt with appropriately for testosterone analyses. Data should have been z-scored within-gender and then statistical tests conducted.

http://faculty.haas.berkeley.edu/dana_carney/pdf_My_position_on_power_poses.pdf


Paper as
Published
Supplemental
Data
@redpenblackpen
Actual work
performed

Researcher Degrees of Freedom

Safeguard: The 21-word solution

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Available at SSRN: https://ssrn.com/abstract=2160588*.

Item	Self-admission rate (%)	
	Control group	BTS group
1. In a paper, failing to report all of a study's dependent measures	63.4	66.5
2. Deciding whether to collect more data after looking to see whether the results were significant	55.9	58.0
3. In a paper, failing to report all of a study's conditions	27.7	27.4
4. Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6	22.5
5. In a paper, "rounding off" a *p* value (e.g., reporting that a *p* value of .054 is less than .05)	22.0	23.3
6. In a paper, selectively reporting studies that "worked"	45.8	50.0
7. Deciding whether to exclude data after looking at the impact of doing so on the results	38.2	43.4
8. In a paper, reporting an unexpected finding as having been predicted from the start	27.0	35.0
9. In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)	3.0	4.5
10. Falsifying data	0.6	1.7

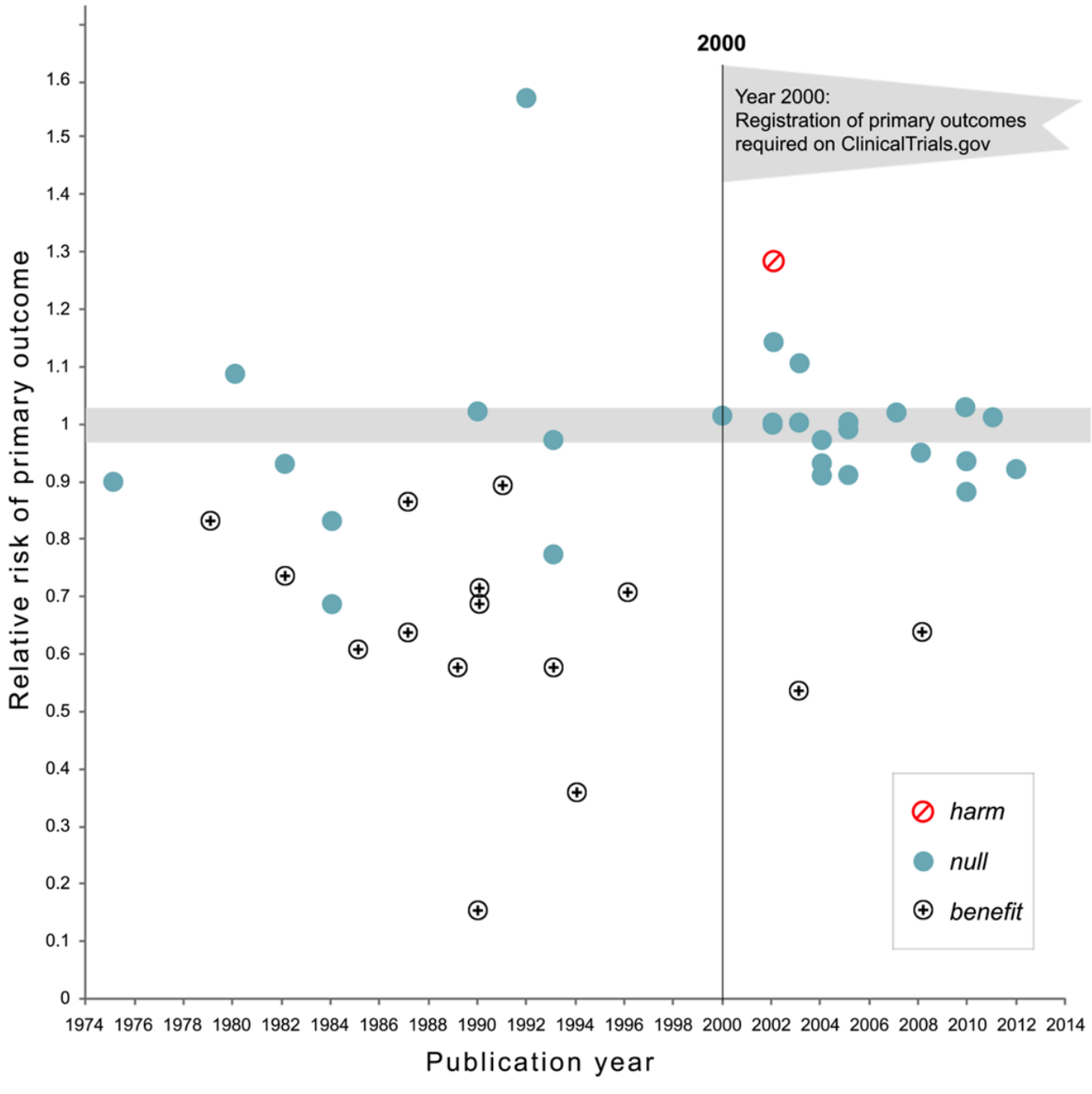
John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Item	Self-admission rate (%)		
	Control group	BTS group	
1. In a paper, failing to report all of a study's dependent measures	63.4	66.5	47.9
2. Deciding whether to collect more data after looking to see whether the results were significant	55.9	58.0	53.2
3. In a paper, failing to report all of a study's conditions	27.7	27.4	16.4
4. Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6	22.5	10.4
5. In a paper, "rounding off" a *p* value (e.g., reporting that a *p* value of .054 is less than .05)	22.0	23.3	22.2
6. In a paper, selectively reporting studies that "worked"	45.8	50.0	40.1
7. Deciding whether to exclude data after looking at the impact of doing so on the results	38.2	43.4	39.7
8. In a paper, reporting an unexpected finding as having been predicted from the start	27.0	35.0	37.4
9. In a paper, claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)	3.0	4.5	3.1
10. Falsifying data	0.6	1.7	2.3

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS ONE, 12*(3), e0172792.

Preregistration makes effects go away

before:
57% success



after:
8% success

Kaplan, R. M. & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One, 10,* e0132382.

Statistical Errors

Revising his paper, the grad student is unable to replicate his own statistical results.

https://legogradstudent.tumblr.com/post/147706279786/revising-his-paper-the-grad-student-is-unable-to

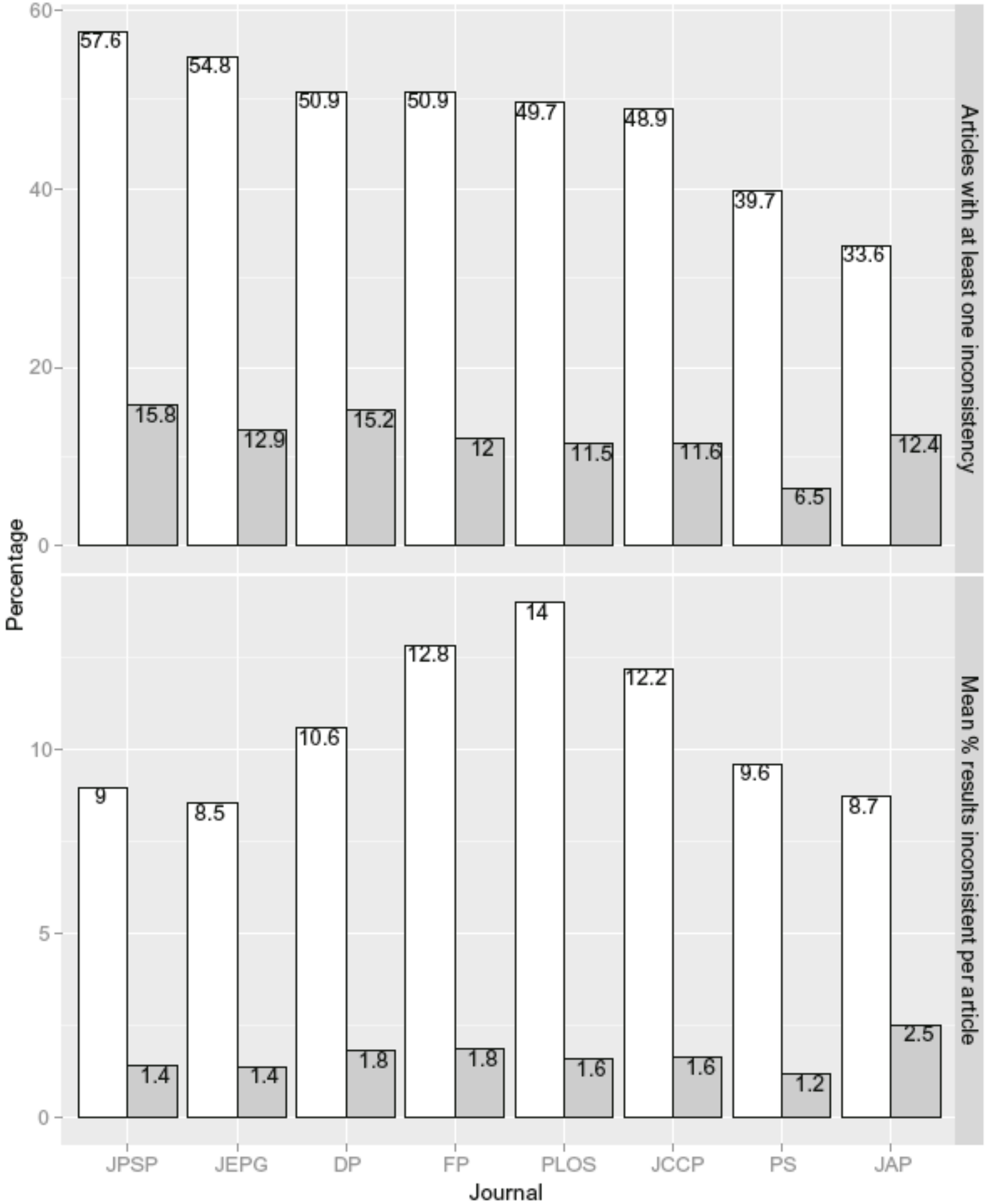
Statistical Errors

statcheck

http://statcheck.io

$$t(118) = 2.45,\ p = .016$$

df + test statistic + p-value ← consistent?



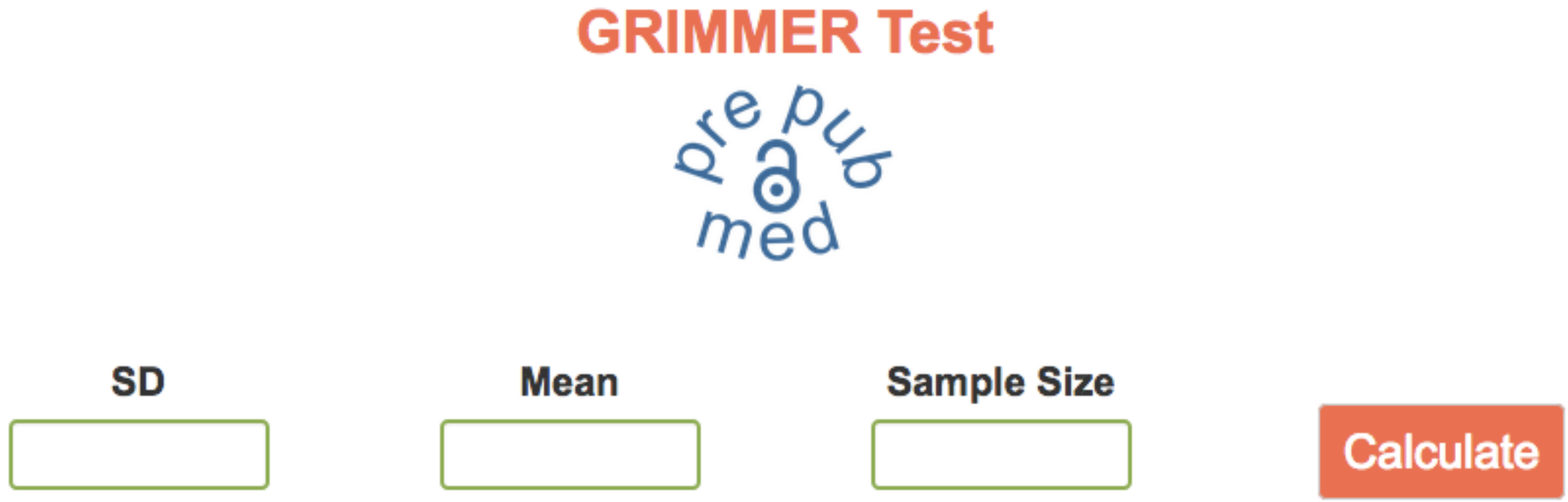
Of all papers reporting NHSTs:

49.6% ≤ 1 inconsistency

12.9% ≤ 1 gross inconsistency

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research, 48*, 1205-1226.

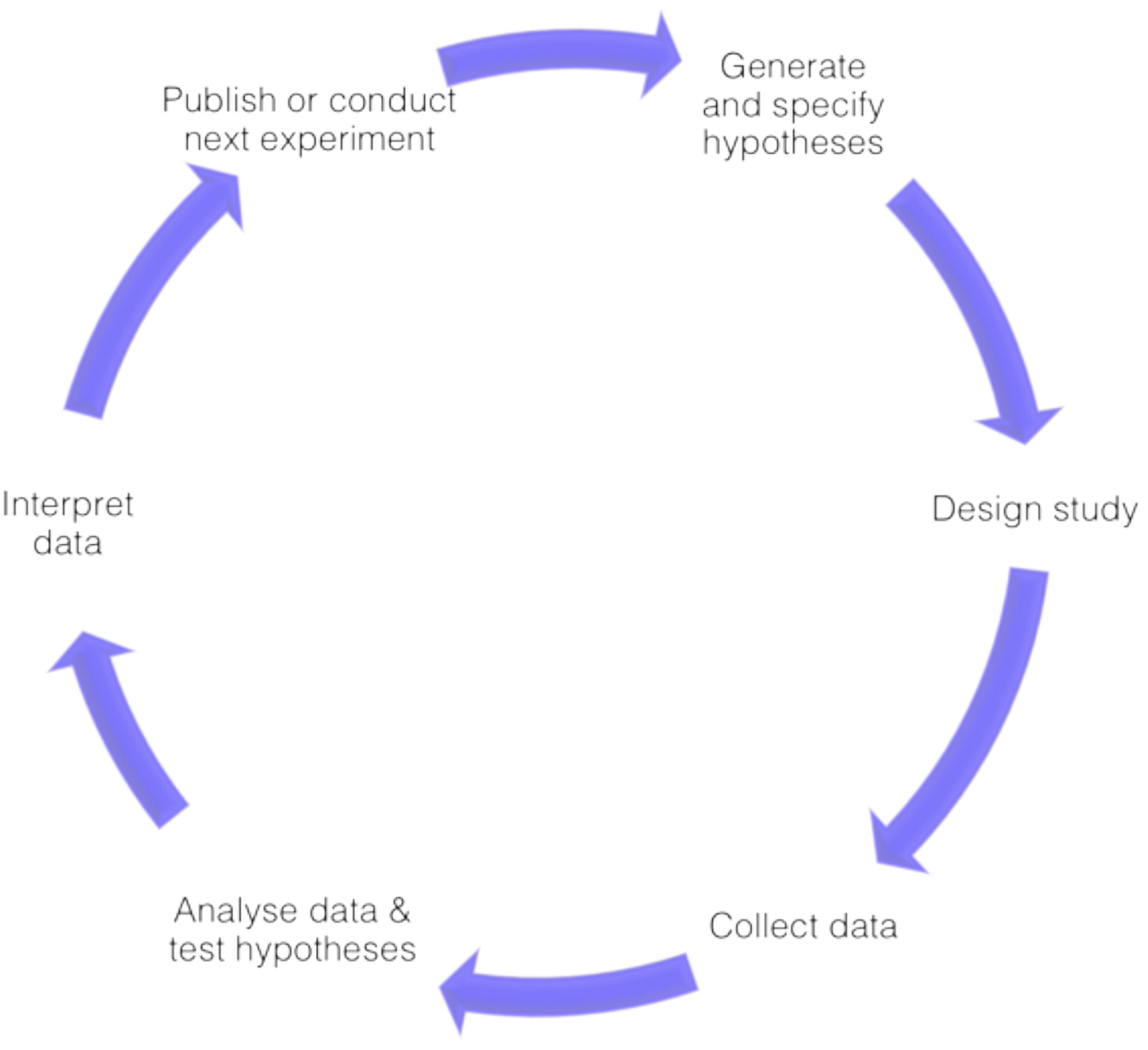
Statistical Errors



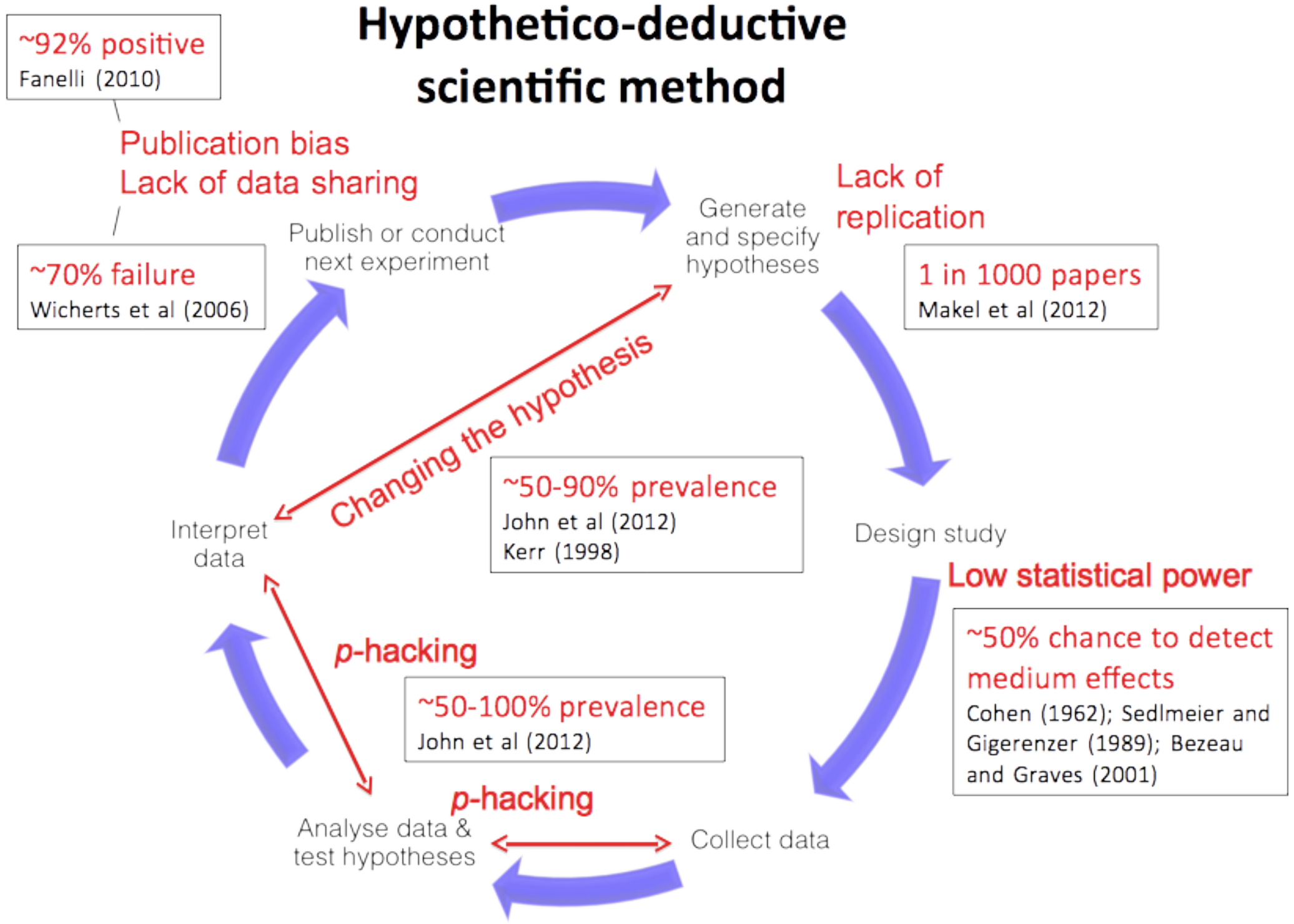
http://www.prepubmed.org/grimmer/

Exercise 3: statcheck & GRIMMER

Hypothetico-deductive scientific method



credit Chris Chambers



credit Chris Chambers

Roots of the Replication Crisis

- Low power
- Publication bias
- Statistical misconceptions
- Researcher degrees of freedom
- Statistical errors
- Lack of replication

"Conventional" psychology research is like cooking a meal from a recipe but arbitrarily adding ingredients and changing procedures until you like the result, *without taking notes*

The Case for Open Science

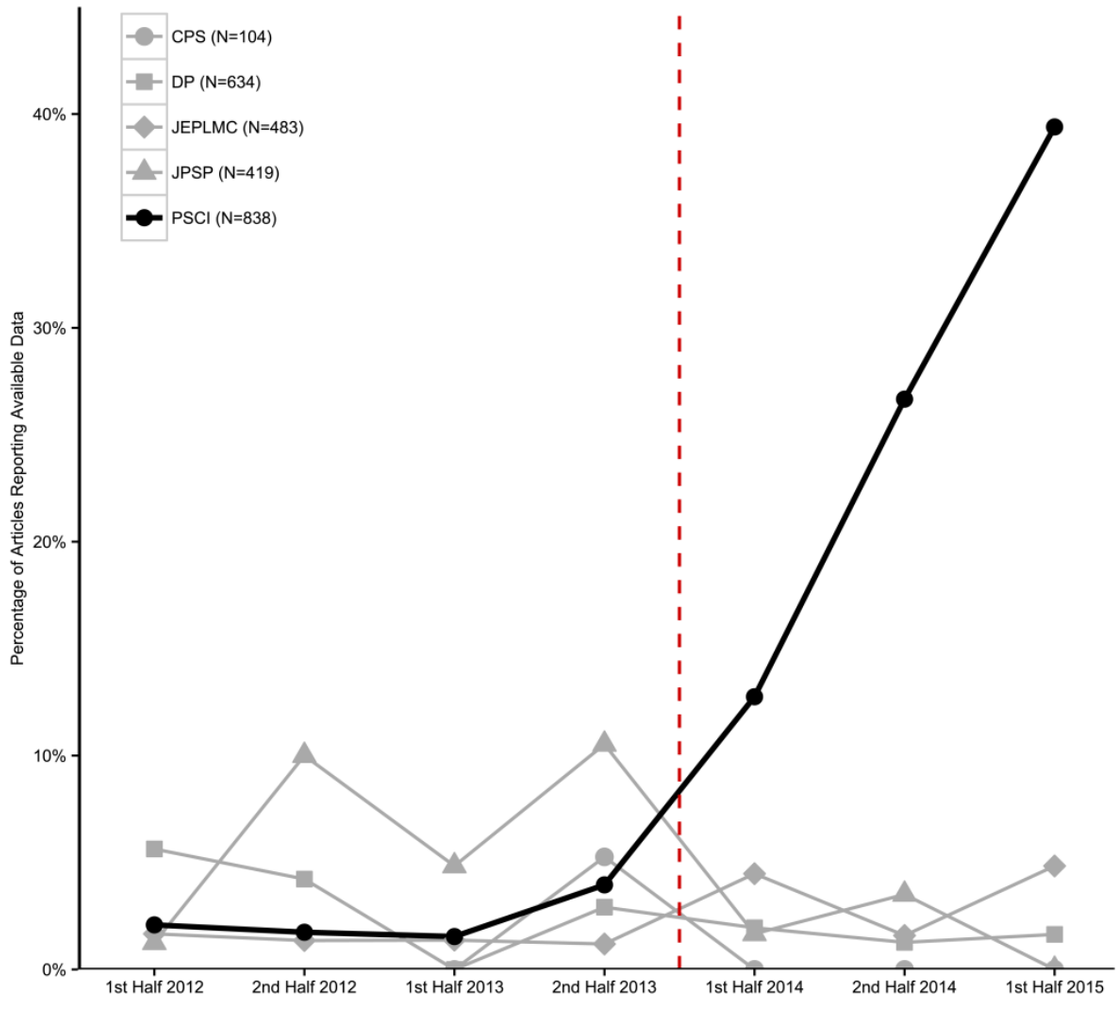
- fix the research cycle → do actual *science*
- accumulate knowledge more effectively and efficiently
- stop getting lost in the sea of researcher degrees of freedom
- stop wasting time trying to understand your own code/analysis/study from X months/years ago
- stop wasting taxpayers' money
- bigger, better studies → more impact
- be prepared for new requirements & standards from journals, funders, employers

→ **future-proof your research**

II. Open Data

Open Data

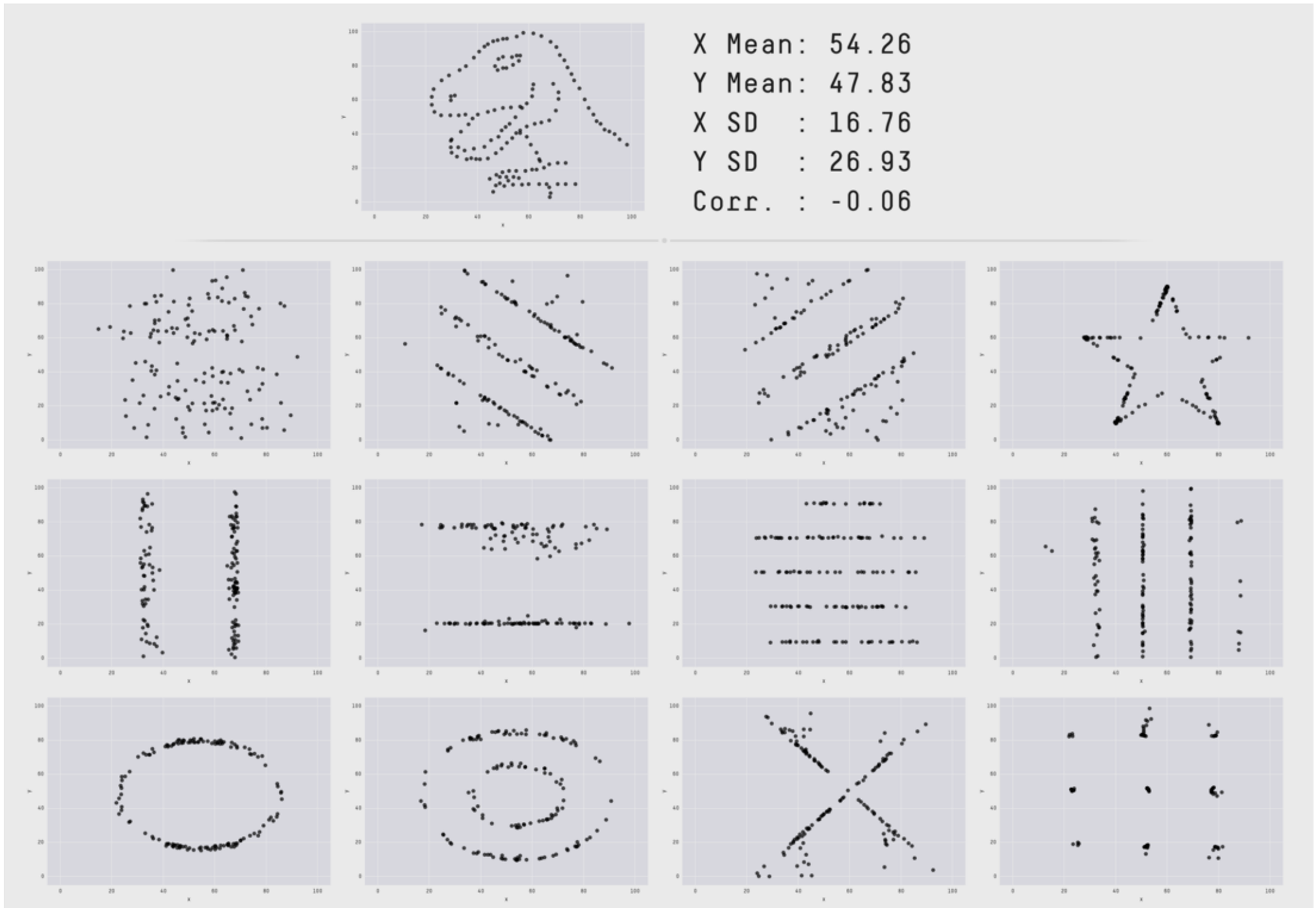

OPEN DATA


CPS (N=104)
DP (N=634)
JEPLMC (N=483)
JPSP (N=419)
PSCI (N=838)
40%
30%
20%
10%
0%
Percentage of Articles Reporting Available Data
1st Half 2012
2nd Half 2012
1st Half 2013
2nd Half 2013
1st Half 2014
2nd Half 2014
1st Half 2015

Open Data: why?

https://www.autodeskresearch.com/publications/samestats

Open Data: why?



Open Data: why?

In all such fields, we can distill Claerbout's insight into a slogan:

An article about computational science in a scientific publication is **not** *the scholarship itself, it is merely* **advertising** *of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.*

Buckheit, J. B., & Donoho, D. L. (1995). *Wavelab and reproducible research*. New York: Springer.

Open Data: why?

German Research Foundation **DFG:**

If data have been generated through publicly funded projects, they are on principle openly available to the public.

Enables secondary data use

→ more efficient use of existing data sets

→ waste less time, money, lab animals; decrease risk for subjects

Mandatory Open Data

↳ **justified exceptions are *always* possible!**

Journals

TOP guidelines

Funding agencies

UK Research Councils
National Science Foundation (USA)
DFG (German Research Foundation)

Employers

More and more positions advertised with open science statements

Before sharing your data

Data privacy concerns

- de-identification
- partial data sharing
- add noise
- share data with restricted access

→ **adapt consent forms!**

Excellent guide to data sharing and consent forms:
https://github.com/gilmore-lab/data-sharing-permission-template

Repositories: Things to Consider

DFG data management guidelines

- economic and ideological independence of the institution
- data should be safe for 10+ years
- what happens with the data if repository stops existing?
- publicly and freely (!) accesssible
- possibility to restrict access
- persistent identifier (DOI or URL)
- usage of repository should not imply giving up rights of use
- option to save data publicly and non-publicly

see also www.datasealofapproval.org/en/

Repositories: Some Options

OSF: http://osf.io/

Germany: PsychData http://psychdata.zpid.de/

Germany: GESIS https://datorium.gesis.org/xmlui

Minimal Folder Structure

dataset

codebook

readme

Exercise 4:

Creating a codebook

Exercise 5:
Using the OSF

III. Preregistration

Preregistration

reduces "questionable research practices"

Registered Reports

http://cos.io/rr

reduces questionable research practices **and** publication bias

Preregistration

↓

confirmatory analyses

planned a priori

hypothesis-driven

controlled error rates (NHST)

exploratory analyses

determined post hoc

data-driven

inflated false discovery rate (capitalising on chance)

p-values meaningless for error control

What's a preregistration?

Stating your...

hypotheses

method

sampling & analysis plan

...*before* collecting the data.

Prevents...

← HARKing

← outcome switching

← *p*-hacking

+ catch design flaws early

+ increase reproducibility

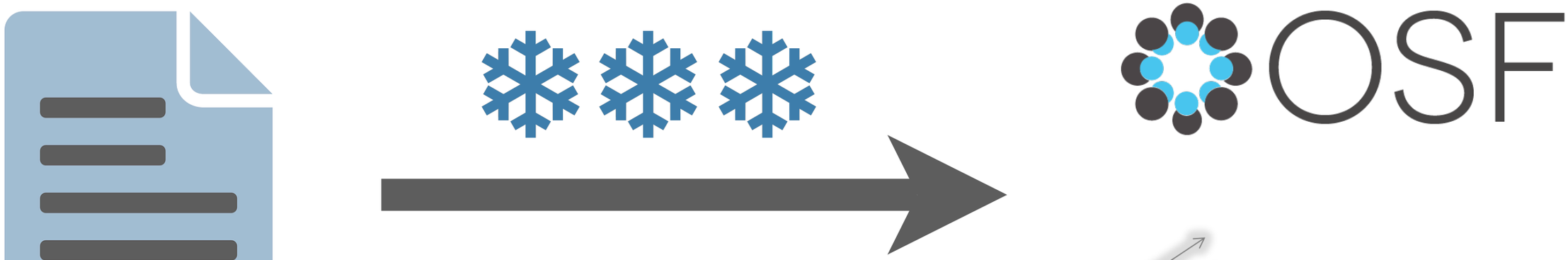
+ salvage null results

Why preregister?

Intervention behavior. On average, infants pressed the button $M = 27.85$ times ($SE = 2.36$) during the test phase. The data of three infants were classified as outliers for exceeding the total mean by more than two standard deviations. They were excluded from further analysis, leaving a final sample of $N = 37$ ($n_{\text{causal chain}} = 12$, $n_{\text{common cause}} = 12$,

The infant's looking was timed starting when the actor's hand or the rod made contact with the toy and continuing until the infant had looked away from the display for 2 s or until 120 s had elapsed. Thus, looking was timed as the baby saw the static

How does it work?



https://osf.io

OSF

https://aspredicted.org

Templates:
osf.io/zab38/wiki/home

Templates: brief

https://aspredicted.org

1) Have any data been collected for this study already?

2) Hypothesis. What's the main question being asked or hypothesis being tested in this study?

3) Dependent variable. Describe the key dependent variable(s) specifying how they will be measured.

4) Conditions. How many and which conditions will participants be assigned to?

5) Analyses. Specify exactly which analyses you will conduct to examine the main question/hypothesis.

6) More analyses. Any secondary analyses?

7) Sample Size. How many observations will be collected or what will determine sample size?

8) Other. Anything else you would like to pre-register?

9) Name. Give a title for this AsPredicted pre-registration

Templates: detailed

Section	Essential elements	Recommended elements
A. Hypotheses	1. Describe the (numbered) hypotheses in terms of relationships between your variables. 2. For interaction effects, describe the expected shape of the interactions. 3. If you are manipulating a variable, make predictions for successful check variables or explain why no manipulation check is included.	4. A figure or table may be helpful to describe complex interactions. 5. For original research, add rationales or theoretical frameworks for why a certain hypothesis is tested. 6. If multiple predictions can be made for the same IV-DV combination, describe what outcome would be predicted by which theory.
B. Method		
Design	List, based on your hypotheses from section A: 1. Independent variables and all their levels a. whether they are within- or between-participants; b. the relationship between them (e.g., orthogonal, nested). 2. Dependent variables. 3. Third variables acting as covariates or moderators.	
Planned sample	4. If applicable, describe pre-selection rules. 5. Indicate where, from whom and how the data will be collected. 6. Justify planned sample size. 7. Describe data collection termination rule.	
Exclusion criteria	8. Describe anticipated data exclusion criteria. Some examples of exclusion criteria are: a. missing, erroneous, or overly consistent responses; b. failing check-tests or suspicion probes; c. demographic exclusions; d. data-based outlier criteria; e. method-based outlier criteria (e.g. too short or long response times).	9. Set fail-safe levels of exclusion at which the whole study needs to be stopped, altered, and restarted.
Procedure	10. Describe all manipulations, measures, materials and procedures including the order of presentation and the method of randomization and blinding (e.g., single or double blind), as in a published Methods section.	
C. Analysis plan		
Confirmatory analyses	Describe the analyses that will test each main prediction from the hypotheses section. For each one, include: 1. the relevant variables and how they are calculated; 2. the statistical technique; 3. each variable's role in the technique (e.g., IV, DV, moderator, mediator, covariate); 4. rationale for each covariate to be used, if any; 5. if using techniques other than null hypothesis testing (for example, Bayesian statistics), describe your criteria and inputs towards making an evidential conclusion, including prior values or distributions.	Specify contingencies and assumptions, such as: 6. method of correction for multiple tests; 7. the method of missing data handling (e.g., pairwise or listwise deletion, imputation, interpolation); 8. reliability criteria for item inclusion in scale; 9. anticipated data transformations; 10. assumptions of analyses, and plans for alternative/corrected analyses if each assumption is violated.

van 't Veer & Giner-Sorolla (2016)
Pre-registration in social psychology: a suggested template

Level of detail

brief

e.g. AsPredicted

\+ low effort, quick

\+ for low-stakes research

\+ exploratory research

\- relatively vague → more RDF

\- inflexible: can't deal with unforeseen circumstances

detailed

e.g. van 't Veer & Giner-Sorolla

relatively strict → fewer RDF +

flexible: sampling safety nets, data-dependent decisions +

for high-stakes research +

more effort -

Pro Tips

- data-dependent decisions (decision trees)
- trial analyses on pilot or simulated data
- **data-collection safety net**
- take the reviewers' perspective
- number everything
  → cite numbers in final report
- don't worry, things never work out perfectly

Prereg for Secondary Analyses

	Nine recommendations to increase transparency in analysis of pre-existing data
1	Provide links to codebooks and data access instructions.
2	Provide the data access form or email.
3	List the specific variables used in the analyses.
4	Post the analytic script prior to analysis.
5	Describe prior by others research using the same variables in the same data set.
6	Describe any analyses, published or unpublished, that you have done on any of the variables used in the current study.
7	Register analysis plan prior to examining data.
8	Preregister analyses for upcoming and not-yet released waves of data.
9	Post analytic script prior to analysis.

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2018, July 6). Recommendations for increasing the transparency of analysis of pre-existing datasets. https://doi.org/10.31234/osf.io/zmt3q

Earning the badge

1) A public, date-time stamped registration is in an institutional registration system.
2) Registration pre-dates the intervention.
3) Registered design and analysis plan corresponds directly to reported design and analysis.
4) Full disclosure of results in accordance with registered plan.

Journals offering the badge

Selection (40 in total):

Journal of Cognition and Development

Language Learning

Psych Science

JESP

BMJ Open Science

Frequent Objections

Preregistration doesn't prevent fraud

sad but true

Preregistration stifles exploration

Preregistration de-stigmatises exploration!

Preregistration doesn't apply to my research

if you test hypotheses, you should preregister

Preregistration is too much extra work

- Preregistration is mostly shifted work, not extra work
- The extra work can dramatically improve your research

Preregistration leads to scooping

- Preregistration gives you proof that you had an idea first
- Preregistrations on the OSF can be embargoed for up to 4 years
- Search for existing preregistrations before starting a study: osf.io/registries

Preregistration makes my effects go away

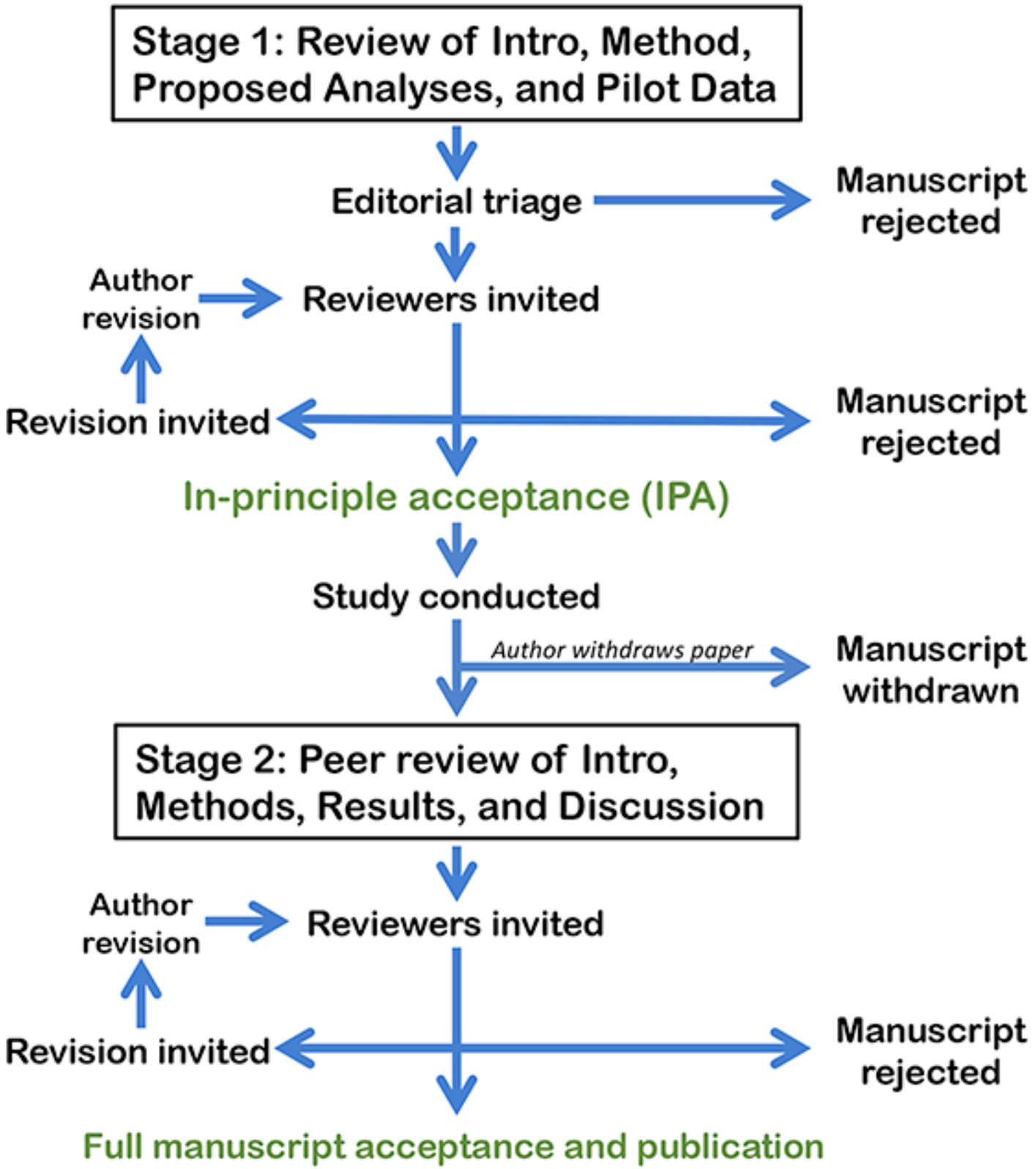
If we do research to find out things about the world,
we should be interested in that!

Exercise 6: Preregistration on the OSF

Registered Reports

http://cos.io/rr

121 journals and counting, e.g. *Cortex, Nature Human Behaviour, Royal Society Open Science*



from http://cos.io/rr

RRs: Stage 1 Review Criteria

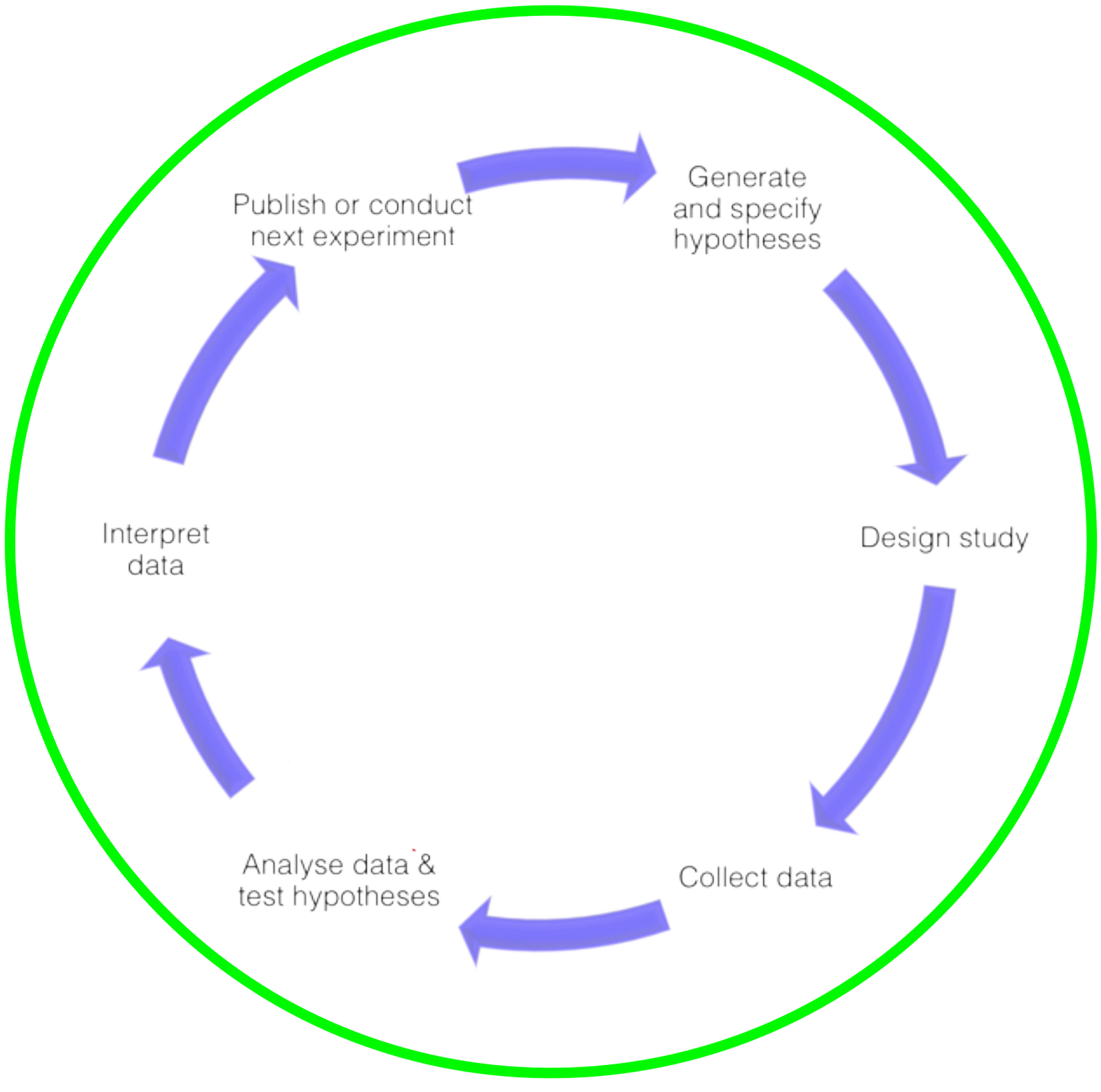
1) The importance of the research question(s).
2) The logic, rationale, and plausibility of the proposed hypotheses.
3) The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis where appropriate).
4) Whether the clarity and degree of methodological detail is sufficient to exactly replicate the proposed experimental procedures and analysis pipeline.
5) Whether the authors have pre-specified sufficient outcome-neutral tests for ensuring that the results obtained are able to test the stated hypotheses, including positive controls and quality checks.

RRs: Stage 2 Review Criteria

1) Whether the data are able to test the authors' proposed hypotheses by satisfying the approved outcome-neutral conditions (such as quality checks, positive controls)
2) Whether the Introduction, rationale and stated hypotheses are the same as the approved Stage 1 submission (required)
3) Whether the authors adhered precisely to the registered experimental procedures
4) Whether any unregistered post hoc analyses added by the authors are justified, methodologically sound, and informative
5) Whether the authors' conclusions are justified given the data

Not relevant:

- $p < .05$
- Whether hypotheses were supported
- Whether the results are novel/interesting/have impact
- Whether the results support Reviewer 2's pet theory



credit Chris Chambers

Registered Reports FAQ

- How long does it take?
  Chris Chambers: Stage 1 review at Cortex takes 9 weeks on average
- What's the acceptance rate?
  Chris Chambers: Normal submissions: ~10%; RRs (after passing editorial triage): 90%; RRs at Stage 2: so far 100%
- What if the method / analysis needs to be changed after Stage 1 review?
  Minor changes can be incorporated (transparently), bigger changes may lead to withdrawal and re-review
- Are RRs possible for existing data?
  Depends: possible at some journals; advantageous if access to data doesn't exist yet
- Are RRs suitable for PhD students?
  Maybe not if data collection needs to start quickly; otherwise: yes, absolutely (basically guaranteed publication after in-principle acceptance, high acceptance rate)

Conclusion

Problems	**Fixes**
Researcher df	← preregistration
Publication bias	← Registered Reports
Low power	← team up, e.g. StudySwap
Errors	← statcheck, reproducible code, **open data**

Thank you!

a.m.scheel@tue.nl

@AnneMScheel

http://tue.nl/staff/a.m.scheel

Slides: https://osf.io/57uyx/