

Unconscious bias in scientific research

Some cautionary tales from Psychology's
“Replication Crisis”

Stuart J. Ritchie

Social, Genetic & Developmental Psychiatry Centre

King's College London

stuart.j.ritchie@kcl.ac.uk

@StuartJRitchie

Not focusing on conscious fraud



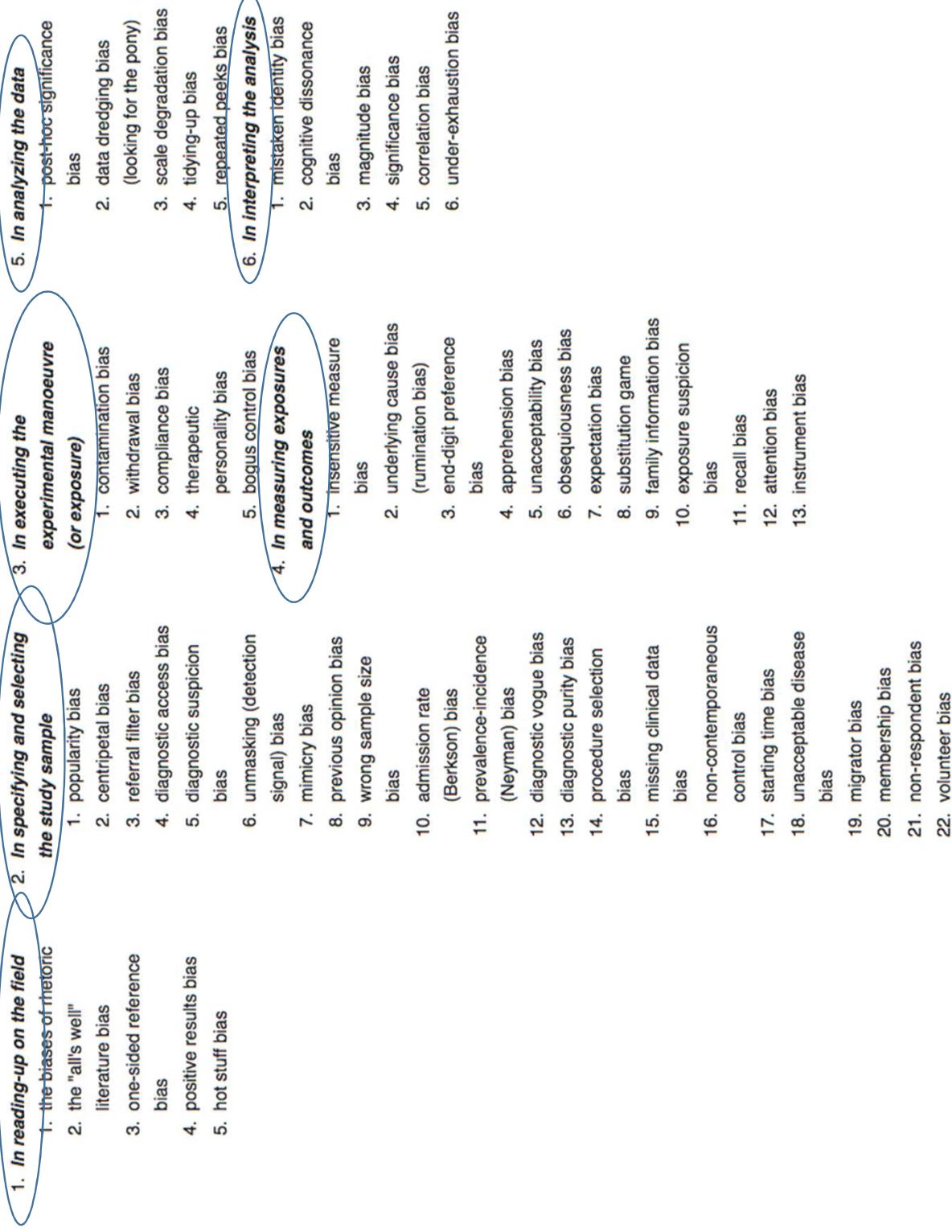
Diederik Stapel

- Tilburg University, NL
- Huge media coverage
- Total data fabrication
- Final(?) number: **58** retracted papers
- Sentenced to community service
- **Nobody ever tried to replicate his studies**

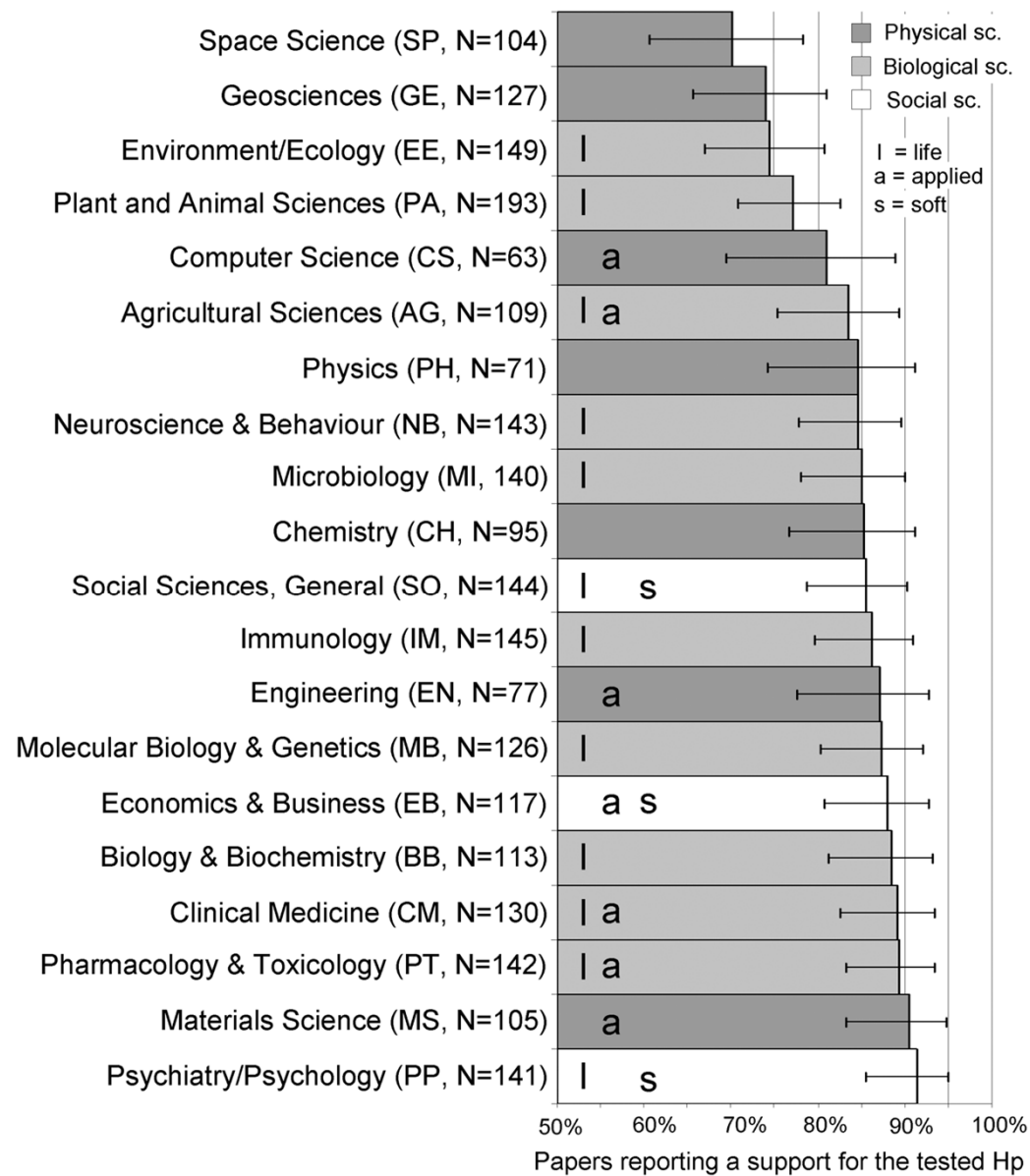
J Chron Dis Vol. 32, pp. 51 to 63
Pergamon Press Ltd 1979. Printed in Great Britain

BIAS IN ANALYTIC RESEARCH

DAVID L. SACKETT



1. Biases in **reading up** on the field



Time-travelling pornography

- 9 experiments
- Cornell undergraduates display psychic powers
 - Sensing pornography
 - Avoiding violent pictures
 - ‘Remembering’ words they’re *about* to see
- Published in a top psychology journal: *JPSP*
- We tried to replicate it...



Bem (2011), *Journal of Personality and Social Psychology*, 100, 407-425

Email from the journal

Dear Dr. Ritchie:

I have read your paper "Failing the future...", submitted to JPSP: Attitudes and Social Cognition as #2011-0072. I found the paper well-written and the findings interesting. Nevertheless, I am writing to inform you that I do not believe that the paper is suitable for this journal, and I must therefore decline it. To save you the time that an extensive review process would take, I am making this decision myself, without the involvement of external reviewers.

The basic issue is this: This journal does not publish replication studies, whether successful or unsuccessful; instead, we seek to publish papers that make a substantive novel theoretical contribution. Although the Bem paper is unusual in many ways, I see no reason to depart from this long-standing journal policy. Of course, a paper that reports a replication as part of a larger study might well make a new theoretical contribution and be publishable here - for example, a paper that includes an exact replication

Eventually published as: Ritchie et al. (2012) *PLOS ONE*, 7, e33423.

2. Biases in **specifying** and **selecting** the sample

Statistical power

- How many subjects do we need, to have 80% power to detect certain effects?
- MTurk (online) sample ($n = 697$)
- Some obvious(?) effect sizes

Can detect with 20 subjects per condition

- Men are taller than women
 - ($n = 6$ per condition)
- People above the average sample age are closer to retirement
 - ($n = 9$ per condition)
- Women own more shoes than men
 - ($n = 15$ per condition)

Cannot detect with 20 subjects per condition

- People who like spicy food more likely to like Indian food
 - (need $n = 26$ per condition)
- Liberals think social equality is more important than do conservatives
 - (need $n = 34$ per condition)
- Men weigh more than women
 - (need $n = 46$ per condition)
- People who like eggs eat egg salad more often
 - (need $n = 48$ per condition)

Bottom line

- “Are you studying an effect bigger than:
 - Men weigh more than women?
- If not, use $n > 50$ ”

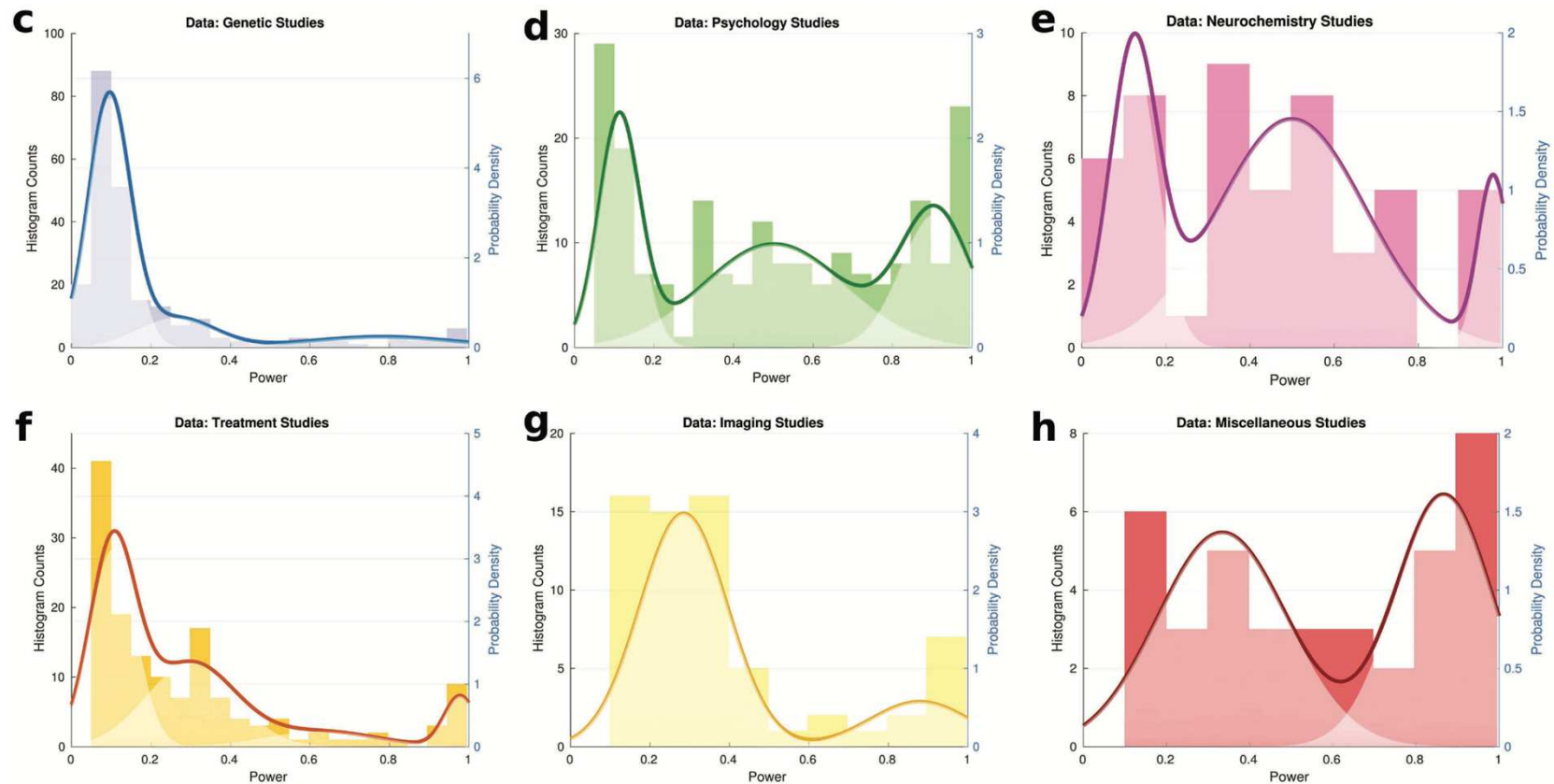
‘Power failure’ in neuroscience

Table 2 | Sample size required to detect sex differences in water maze and radial maze performance

	Total animals used	Required <i>N</i> per study		Typical <i>N</i> per study		Detectable effect for typical <i>N</i>	
		80% power	95% power	Mean	Median	80% power	95% power
Water maze	420	134	220	22	20	$d=1.26$	$d=1.62$
Radial maze	514	68	112	24	20	$d=1.20$	$d=1.54$

Meta-analysis indicated an effect size of Cohen's $d = 0.49$ for water maze studies and $d = 0.69$ for radial maze studies.

Re-analysis by subject area



Nord et al. (2017) *J Neurosci*, 3592-16

3 & 4. Biases in **executing** the
experiment, and **measuring**
exposures and outcomes

Bargh et al. (1996)

- Stereotype priming
 - Cited **3,752** times
 - Participants primed non-consciously
 - Exposed to ageing-related words in a scrambled sentence task
 - Measured: Walking speed
 - Subjects primed by elderly stereotypes walked away from the lab more slowly



Bargh et al. (1996). *Journal of Personality and Social Psychology*, 71(2), 230-244.

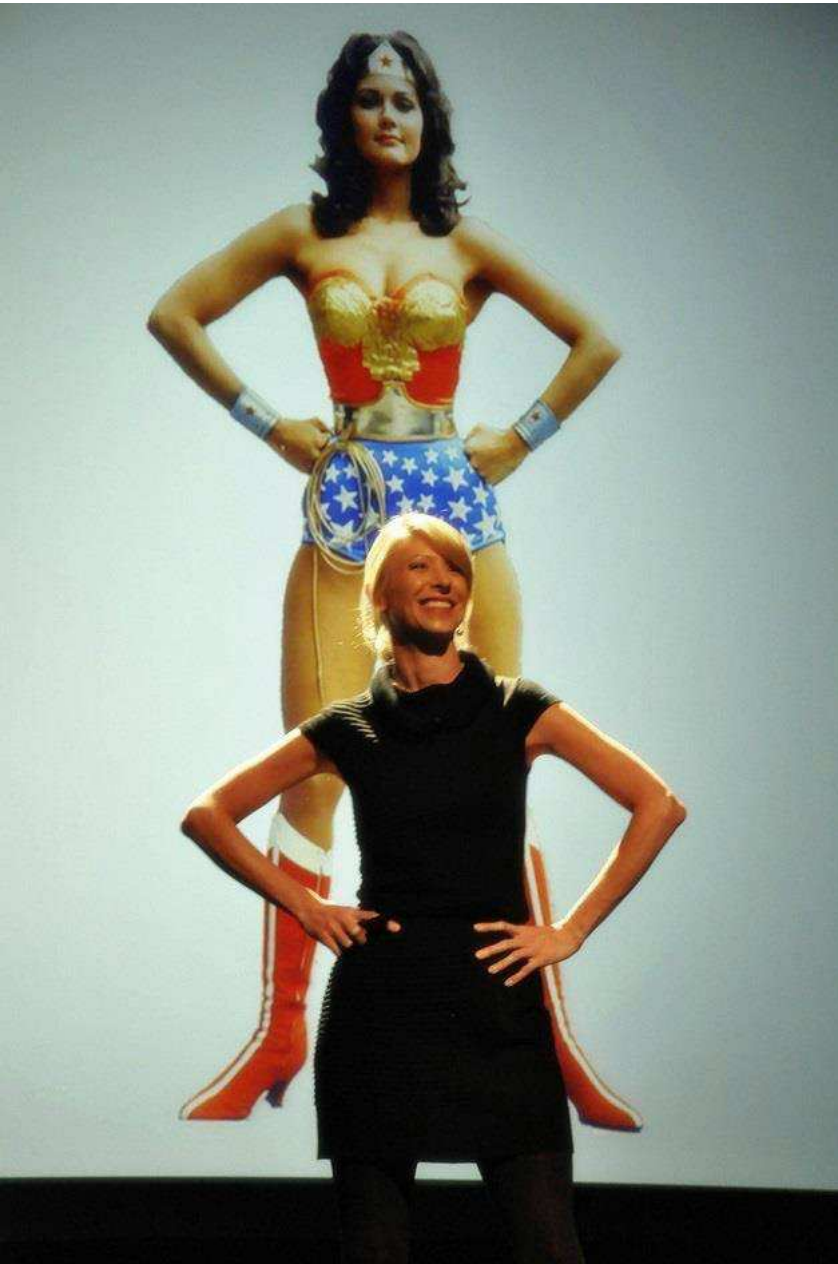
Replication

- Doyen et al. (2012) doubled the number of participants
 - Greater statistical power
- Experiment 1
 - Timed speed with infra-red beams (Bargh et al. had assistant use a stopwatch)
 - Experimenters blind to condition and expectation
 - **Null results**
- Experiment 2
 - Experimenters know the expected result and which condition participants had been allocated to (unblinded)
 - **Slowing effect was observed**

5. Biases in **analysing** the data

‘Researcher degrees of freedom’

- ‘Questionable Research Practices’, A.K.A. p -hacking:
- Optional Stopping / Data-peeking
 - “Let’s just collect a few more subjects...”
- Including/excluding subjects/data/covariates after looking at the results
- Changing analyses after looking at the results
- Failing to report non-significant results
 - Only the ones that ‘worked’



‘Power Posing’

- Carney, Cuddy, & Yap (2010) *Psychol Sci*, 21, 1368-1368
 - Cited **469** times
- Power posing = higher testosterone, lower cortisol, higher feeling of power, higher risk tolerance
- Cuddy’s TED talk: 46.8m views + 14m on YouTube
- Ranehill et al. (2015) *Psychological Science*, 26(5), 653-656
 - Failure to replicate

Co-author's letter on Power Posing

- “As evidence has come in over these past 2+ years, my views have updated to reflect the evidence. *As such, I do not believe that “power pose” effects are real.*”
- Sample size tiny ($n = 42$)
- Initially, outcome of interest was risk-taking
- Checked the significance of the effect along the way – 25 subjects, then added 10, then added 7, then added 5
- 5 exclusions – “didn’t follow directions” – not reported in paper
- Ran multiple statistical tests, picked the one with the lowest p -value
- Outliers dropped from some, but not all, analyses
- Asked many questions, only reported the ones that gave positive results

http://faculty.haas.berkeley.edu/dana_carney/pdf_My%20position%20on%20power%20poses.p

The Cornell Food and Brand Lab

- <https://www.youtube.com/watch?v=8u6xdGCIq6o>
- Huge funding
- Media success
- Influence on policy
- Hundreds of scientific papers
- Wrote a blog in Nov. 2016...



“The grad student who never said ’no’”

- <https://web.archive.org/web/20170312041524/http://www.brianwansink.com/phd-advice/the-grad-student-who-never-said-no>


Every day she came back with puzzling new results, and every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses. Eventually we started discovering solutions that held up regardless of how we pressure-tested them. I outlined the first paper, and she wrote it up, and every day for a month I told her how to rewrite it and she did. This happened with a second paper, and then a third paper (which was one that was based on her own discovery while digging through the data).

RESEARCH ARTICLE

Open Access

Statistical heartburn: an attempt to digest four pizza publications from the Cornell Food and Brand Lab



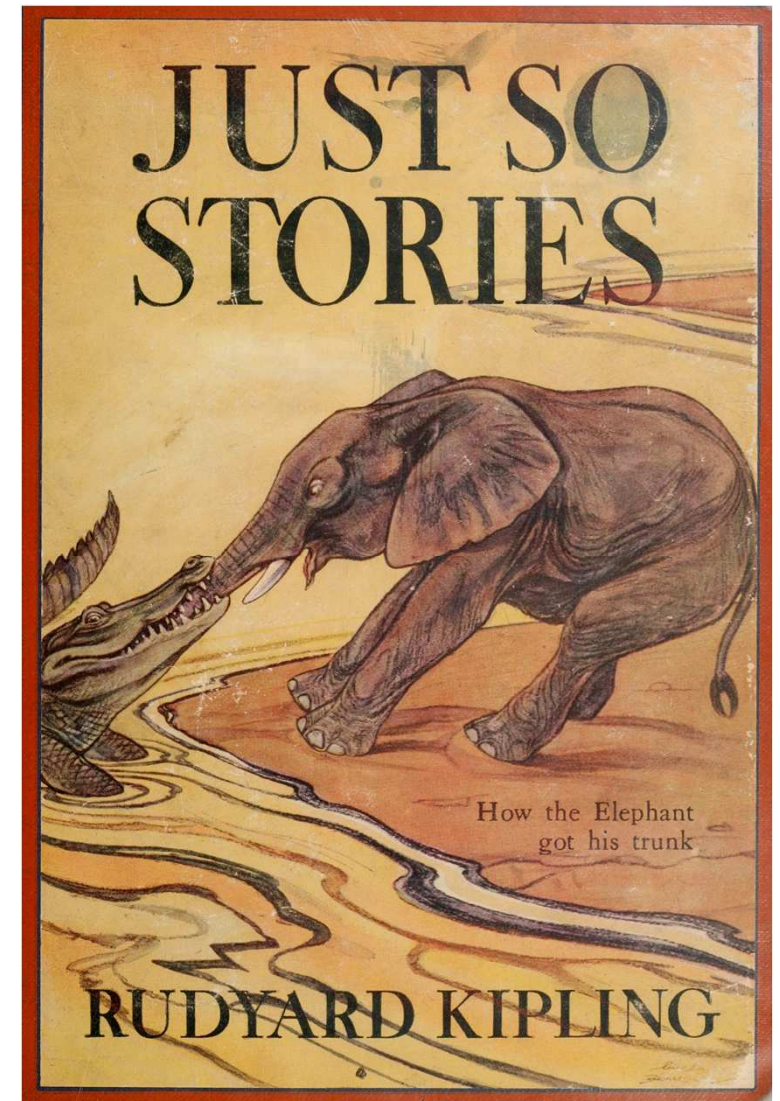
Tim van der Zee¹, Jordan Anaya² and Nicholas J. L. Brown^{3*} 

- 150 errors found across four papers about pizza
- 45 total papers alleged to contain issues (statistical/data inconsistencies, data duplication, self-plagiarism)
- Now 13 retracted papers (and several corrections)
- In October 2018, announced early retirement from Cornell

6. Biases in **interpreting** the analysis

HARKing

- **Hypothesizing After Results are Known**
- Flexibility in which outcome was ‘meant’ to be tested
- Flexibility in the explanation for unexpected results
- See also: CARKing
 - Critiquing After Results are Known



Excuse-making

- Matthew Hankins's 'Still Not Significant'
 - <https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/>

“A trend that approached significance...” ($p < .06$)

“Hovering close to significance” ($p = .076$)

“A trend significance level” ($p = .08$)

“All but significant” ($p = .055$)

“Narrowly eluded statistical significance” ($p = .0789$)

“Barely escapes statistical significance” ($p = .07$)

“Approaching marginal significance” ($p = .064$)

“Fairly significant” ($p = .09$)

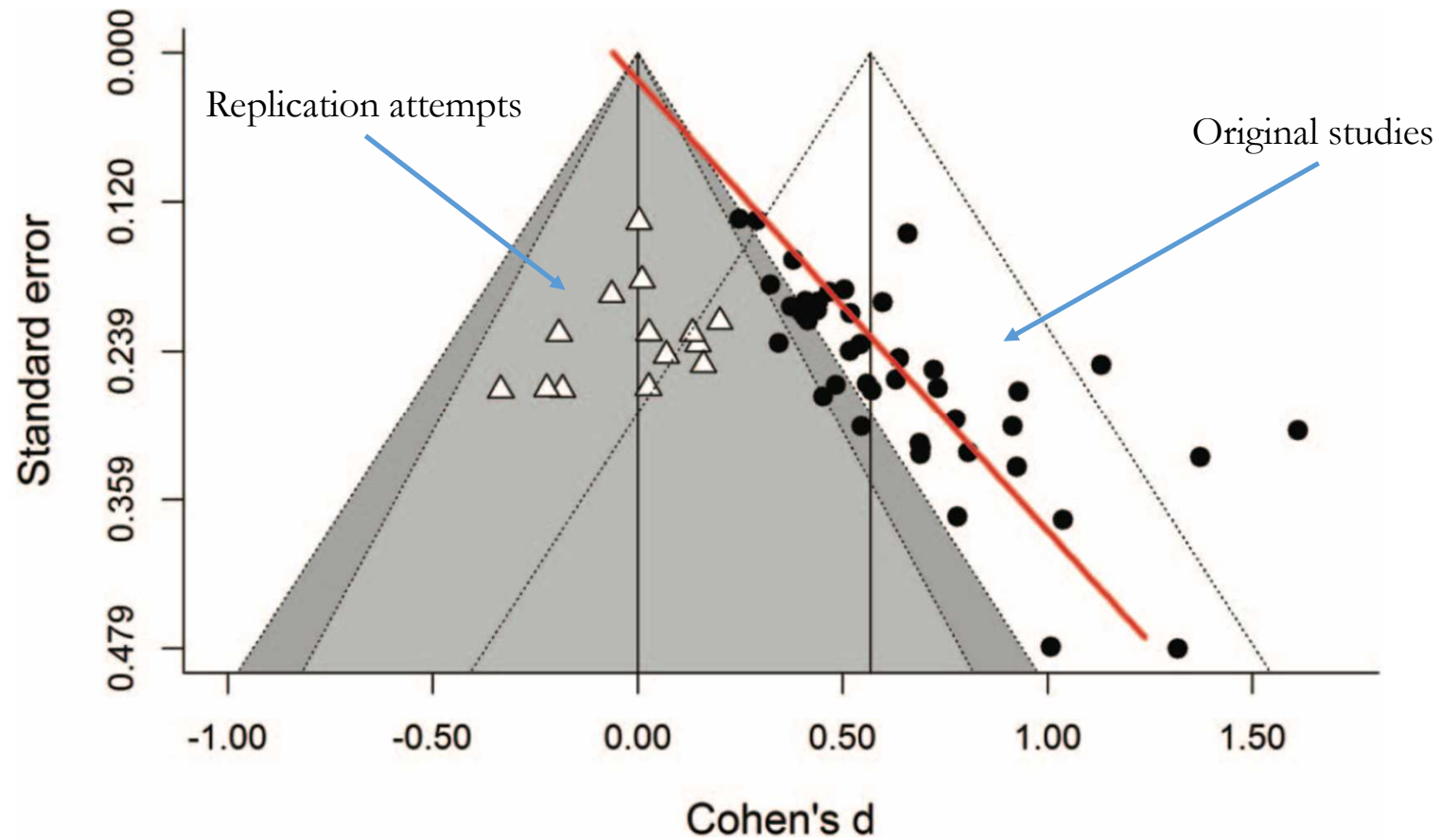
“Very closely brushed the limit of statistical significance” ($p = .051$)

“Significantly significant” ($p = .065$)

“Not absolutely significant but very probably so” ($p > .05$)

7. Full circle: biases in **publishing**
the results

Publication bias



Shanks et al. (2015) *J Exp Psychol Gen* 114, e148-158

“White hat” bias



International Journal of Obesity (2010) 34, 84–88
© 2010 Macmillan Publishers Limited All rights reserved 0307-0565/10 \$32.00

www.nature.com/ijo

COMMENTARY

White hat bias: examples of its presence in obesity research and a call for renewed commitment to faithfulness in research reporting

MB Cope¹ and DB Allison²

How do we **correct** these biases?

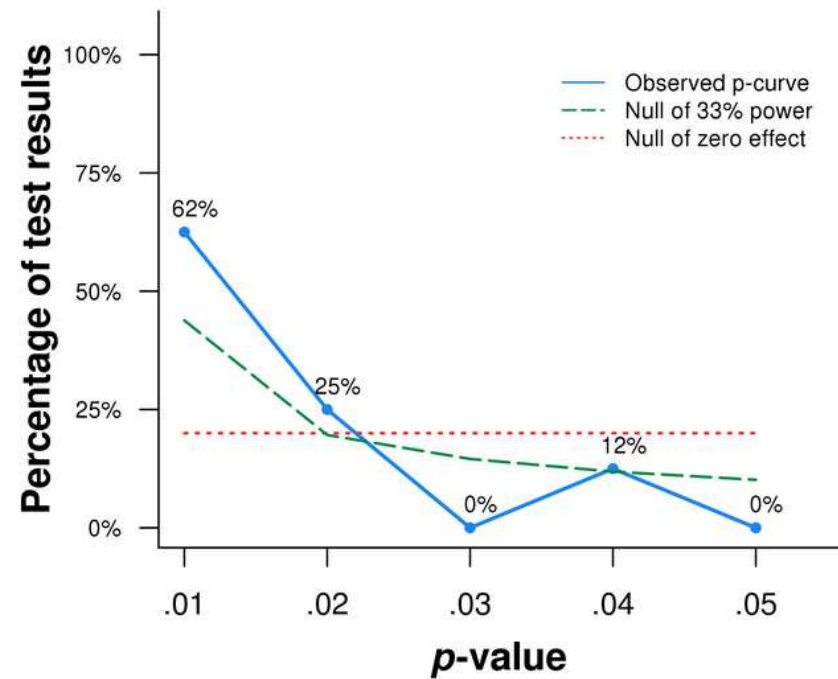
New rules for research

Requirements for authors

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.
2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.
3. Authors must list all variables collected in a study.
4. Authors must report all experimental conditions, including failed manipulations.
5. If observations are eliminated, authors must also report what the statistical results are if those observations are included.
6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

New methods for detection

www.p-curve.com




Statistical Inference	Results	
	Binomial Test <small>(Share of significant results $p < .025$)</small>	Continuous Test <small>(Aggregate pp-values via Stouffer Method)</small>
1) Studies contain evidential value. <small>(Right skew)</small>	$p = .0352$	$Z = -3.94, p < .0001$
2) Studies' evidential value, if any, is inadequate. <small>(Flatter than 33% power)</small>	$p = .9224$	$Z = 1.83, p = .9663$
3) Studies exhibit evidence of intense p -hacking. <small>(Left skew)</small>	$p = .9961$	$Z = 3.94, p > .9999$
Estimate of Statistical Power		
Average power of tests included in p -curve <small>(correcting for publication bias)</small>	71%	


The observed p -curve includes 8 significant results ($p < .05$), of which 87.5% are $p < .025$.
There were no non-significant results entered.

Open data, open materials


- Open Science badges

**Open Data Badge**

- A URL, doi, or other permanent path for accessing the data in a public, open-access repository
- Sufficient information for an independent researcher to reproduce the reported results

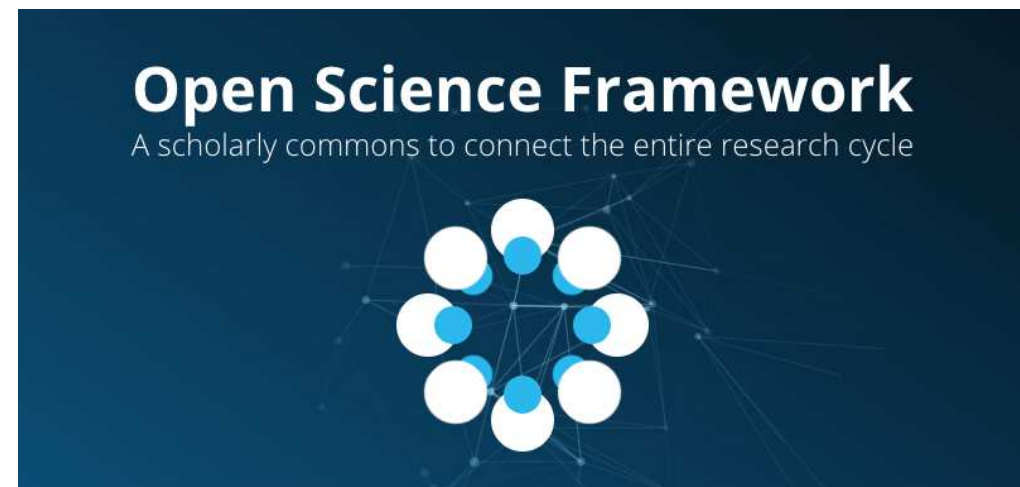
**Open Materials Badge**

- A URL, doi, or other permanent path for accessing the materials in a public, open-access repository
- Sufficient information for an independent researcher to reproduce the reported methodology

**Preregistration Badge***

- URL, doi, or other permanent path to the registration in a public, open-access repository
- An analysis plan registered prior to examination of the data or observing the outcomes
- Any additional registrations for the study other than the one reported
- Any changes to the preregistered analysis plan for the primary confirmatory analysis
- All of the analyses described in the registered plan reported in the article

- Open Science Framework



<http://osf.io>

Transparency and Openness Promotion (TOP) Guidelines for journals

Citation Standards Describes citation of data	Data Transparency Describes availability and sharing of data
Analytical Methods Transparency Describes analytical code accessibility	Research Materials Transparency Describes research materials accessibility
Design and Analysis Transparency Sets standards for research design disclosures	Preregistration of Studies Specification of study details before data collection
Preregistration of Analysis Plans Specification of analytical details before data collection	Replication Encourages publication of replication studies

- <https://cos.io/our-services/top-guidelines/>

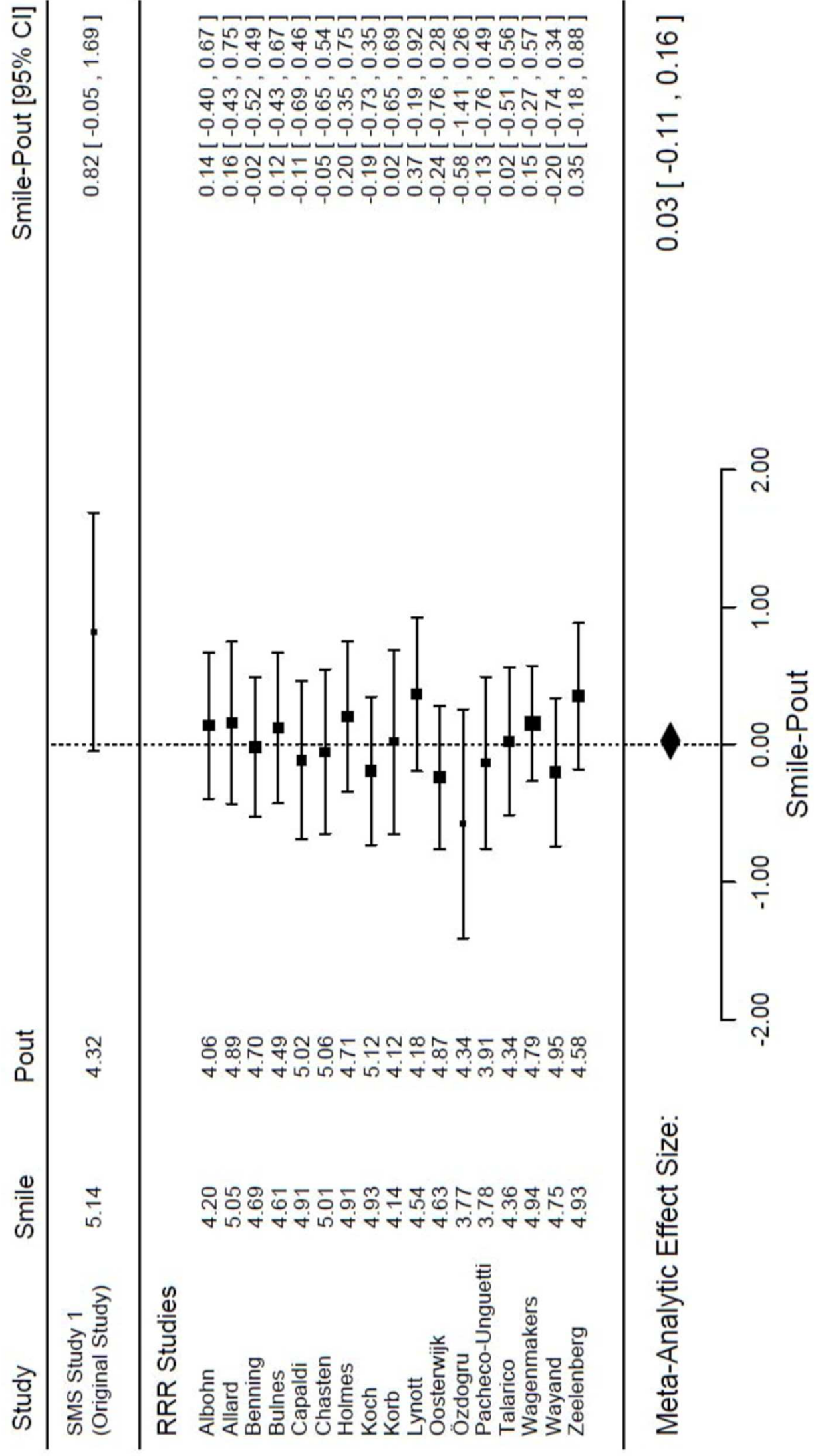
Pre-registration

- ‘Registered Report’: send Intro and Method to be peer-reviewed
- Pre-set statistical analyses
 - Declare any additional ‘exploratory’ analyses
- Provide time-stamped results/analysis files
- Journal accepts the paper *whichever way the results go*
- Deals with:
 - File-drawer bias
 - Methodological biases
 - Researcher degrees of freedom/QRPs
 - HARKing and CARKing
- Not appropriate for all study types/datasets?
 - Can minimally pre-register by posting analysis plan before running analyses
- **Now available at 129 journals! <https://cos.io/rr/>**

Pre-registered replications

- Facial Feedback Hypothesis
 - Strack et al. (1988) *Journal of Personality and Social Psychology*, 54, 768-777
 - Cited **1,590** times
- Registered Replication Report:
 - Acosta et al. (2016) *Perspectives on Psychological Science*, 11, 917-928
 - 17 independent, preregistered replication experiments







Further reading

- Reading list:
 - <http://crystalprisonzone.blogspot.co.uk/2016/03/a-reading-list-for-replicability-crisis.html>
- Summary from 2016:
 - <http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/>
- Alternative (wrong I.M.O.) perspectives:
 - Finkel et al. (2015) *Journal of Personality and Social Psychology*, 108(2), 275-297
 - Gilbert et al. (2016) *Science*, 351(6277), 1037
 - But see: <http://datacolada.org/47>
- Internet comic about psychology's replication crisis:
 - <https://thenib.com/repeat-after-me>

stuart.j.ritchie@kcl.ac.uk

@StuartJRitchie