Open Science, Reproducibility, and Psychology

Andrew Stewart

Division of Neuroscience and Experimental Psychology,

University of Manchester

andrew.stewart@manchester.ac.uk

@ajstewart_lang

Software Sustainability Institute

BETTER SOFTWARE BETTER RESEARCH

www.software.ac.uk

We have a replication problem...

Replication and Reproducibility in Science

- Ionnidis (2005), *PLOS Medicine*, most published research findings are false.
- Prinz et al. (2011), *Nature Reviews Drug Discovery*, around 65% of cancer biology studies do not replicate.
- Button et al. (2013), *Nature Reviews Neuroscience*, small sample size undermines the reliability of neuroscience.
- MacLeod et al. (2014), *Lancet*, 85% of biomedical research resources are wasted.
- Baker (2015), *Nature*, 90% of scientists recognise a ‘reproducibility crisis’.
- Nosek & Errington (2017), *eLife*, out of first 5 replication attempts of preclinical cancer biology work, only 2 have replicated.
- Eisner (2018), *Journal of Molecular and Cellular Cardiology*. Reproducibility of science: Fraud, impact factors and carelessness.

How did we get to where we are?

2011 - 2012

In 2011, the *Journal of Personality and Social Psychology* published a paper by Daryl Bem showing that the future can influence the present - in one study, participants were better able to recall words that they were **later** randomly assigned to rehearse.

This paper used standard statistical methods and ways of doing science.

So, either physics is wrong or the way in which we have been doing science is wrong.

Nelson, L.D, Simmons, J. & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*.

2011 - 2012

Again in 2011, Simmons, Nelson, and Simonsohn published the paper “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant” in *Psychological Science*.

They show that selectively reporting data (e.g., dropping participants, ‘problematic’ trials) and selectively reporting analyses (e.g., only reporting comparisons that are significant) results in vastly inflated false positives.

Later termed *p*-hacking.

Nelson, L.D, Simmons, J. & Simonsohn, U. (2018). Psychology’s Renaissance. *Annual Review of Psychology*.

General Article

Psychological Science
22(11) 1359–1366

DOI: 10.1177/0956797611417632
http://pss.sagepub.com

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

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

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

2011 - 2012

Doyen et al. (2012) failed to replicate the influential Bargh work on social priming - that priming participants with words that activate stereotypes of elderly people results in those participants walking more slowly.

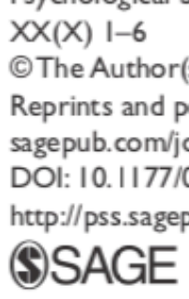
In 2011, Brian Nosek set up replication attempts to try to determine how big a replication issue psychology might be facing. This resulted in the establishment of the Centre for Open Science (2012).

Nelson, L.D, Simmons, J. & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*.

Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance

Psychological Science
XX(X) 1–6

DOI: 10.1177/0956797610383437
http://pss.sagepub.com



Dana R. Carney[1], Amy J.C. Cuddy[2], and Andy J. Yap[1]
[1]Columbia University and [2]Harvard University

Abstract

Humans and other animals express power through open, expansive postures, and they express powerlessness through closed, contractive postures. But can these postures actually cause power? The results of this study confirmed our prediction that posing in high-power nonverbal displays (as opposed to low-power nonverbal displays) would cause neuroendocrine and behavioral changes for both male and female participants: High-power posers experienced elevations in testosterone, decreases in cortisol, and increased feelings of power and tolerance for risk; low-power posers exhibited the opposite pattern. In short, posing in displays of power caused advantaged and adaptive psychological, physiological, and behavioral changes, and these findings suggest that embodiment extends beyond mere thinking and feeling, to physiology and subsequent behavioral choices. That a person can, by assuming two simple 1-min poses, embody power and instantly become more powerful has real-world, actionable implications.


SECURITY
STABILITY
OPPORTU

Power Posing - 2010 *vs.* 2016

Appearance: Big … very big. Spread your hands and legs wide, argued the authors, and you will both exude power and – this was the new finding – feel great. Adopt a power pose and your testosterone rises and your stress levels fall. Or, as columnist David Brooks neatly put it: "If you act powerfully, you will begin to think powerfully."

And now? Well, that's the odd thing. One of the original report's three authors, Dana Carney, says it was all nonsense. "I do not believe that 'power pose' effects are real," she wrote in a blog that detailed the original research's methodological failings. Standing like John Wayne in a gunfight does not make you feel like a successful gunslinger. It just makes you look silly.

https://www.theguardian.com/politics/shortcuts/2016/sep/28/george-osbornes-power-pose-the-science-proves-feeble

Failed replications or effect sizes much smaller than in the original…

- Power posing
- Ego depletion
- Social priming
- Marshmallow test performance predicts future achievement
- Stanford prison experiment
- Growth mindset
- Any others you know of?

https://digest.bps.org.uk/2016/09/16/ten-famous-psychology-findings-that-its-been-difficult-to-replicate/

Why are so many studies not replicating?

- There are too many studies with experimental power too low to detect the effect size of interest.

- One of the consequences of a low powered study is that when real effects are detected their magnitude is likely to be over-estimated.

- Studies which find the effect are published and studies that don’t are not published - due to a bias to publish positive results.

- Future work may use the published effect size during *a priori* power analysis (and then fail to find the effect as the new study is effectively under-powered for what it’s looking for).

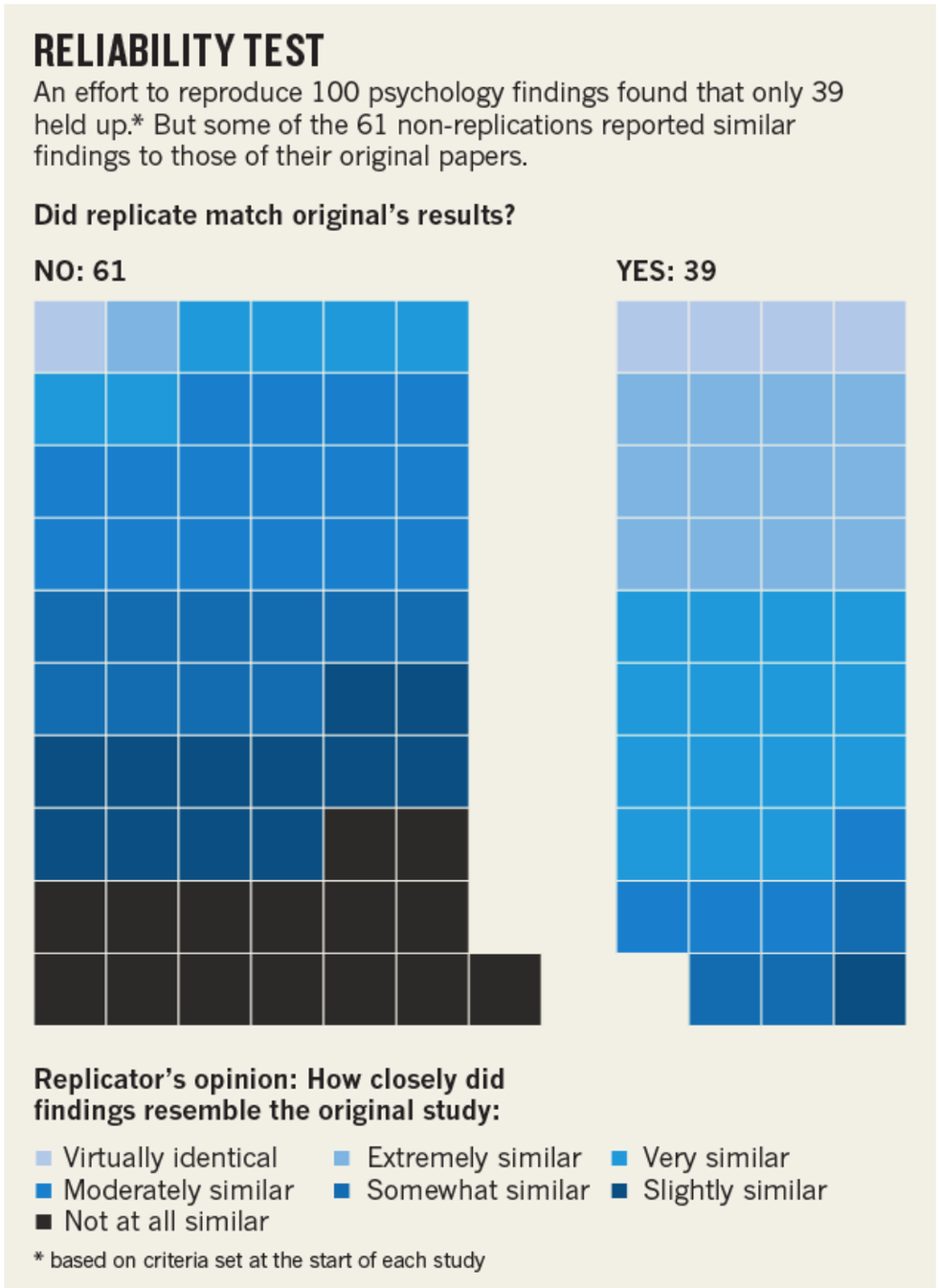
- Button et al. (2013), *Nature Reviews Neuroscience*, small sample size undermines the reliability of neuroscience. Nord et al., (2017), *Journal of Neuroscience,* highlight wide heterogeneity in power in neuroscience studies.

Table 2. Median, maximum, and minimum power subdivided by study type

Group	Median power (%)	Minimum power (%)	Maximum power (%)	2.5th and 97.5th percentile (based on raw data)	95% HDI (based on GMMs)	Total N
All studies	23	0.05	1	0.05–1.00	0.00 – 0.72, 0.80 – 1.00	730
All studies excluding null	30	0.05	1	0.05–1.00	0.01 – 0.73, 0.79 – 1.00	638
Genetic	11	0.05	1	0.05 – 0.94	0.00 – 0.44, 0.63 – 0.93	234
Treatment	20	0.05	1	0.05–1.00	0.00 – 0.65, 0.91–1.00	145
Psychology	50	0.07	1	0.07–1.00	0.02 – 0.24, 0.28 – 1.00	198
Imaging	32	0.11	1	0.11–1.00	0.03 – 0.54, 0.71–1.00	65
Neurochemistry	47	0.07	1	0.07–1.00	0.02 – 0.79, 0.92–1.00	50
Miscellaneous	57	0.11	1	0.11–1.00	0.09 – 1.00	38

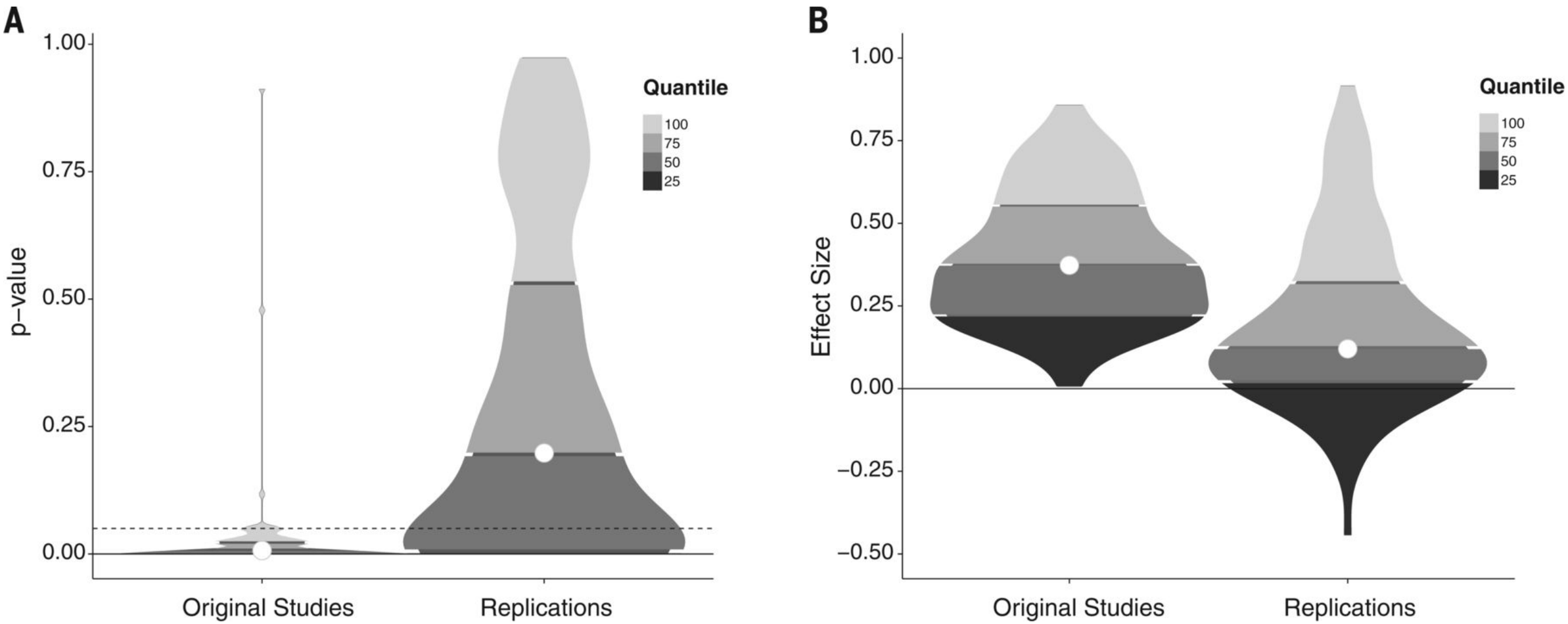
How big an issue is replication for Psychology?

Estimating the reproducibility of psychological science (Nosek et al., 2015)



270 authors tried to replicate 100 experiments drawn from high profile Psychology journals - *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

https://science.sciencemag.org/content/349/6251/aac4716



The *p*-values for the replication set formed a very different distribution to the *p*-values of the original studies. Similarly with the distribution of effect sizes.

https://science.sciencemag.org/content/349/6251/aac4716

2018 - Many Labs 2

186 authors from 36 nations attempted to replicate 28 findings with ~7,000 participants per study.

Average sample size was 64x times the size of the original.

50% of the original studies replicated.

The effect sizes in three quarters of the replications were smaller than the size of the original effect sizes and the effect sizes of 9 of the studies were in the opposite direction to the original!

	Original Study		Replication					
				Global effects		Significance Tests by Sample		
Effect	ES	95% CI	Median ES	ES	95% CI	Percentage <0 (p<.05)	Percentage ns	Percentage >0 (p<.05)
Cohen's q Effect Size								
Disgust & Homophobia (Inbar et al., 2009)	0.70	.05, 1.36	0.03	0.05	.01, .10	3.39	93.22	3.39
Assimilation & Contrast (Schwarz et al., 1991)	0.48	.07, .88	-0.06	-0.07	-.12, -.02	5.08	91.53	3.39
Cohen's d Effect Size								
Correspondence Bias (Miyamoto & Kitayama, 2002) - WEIRD	2.47	1.46, 3.49	1.78	1.81	1.75, 1.88	0.00	0.00	100.00
Correspondence Bias (Miyamoto & Kitayama, 2002) - less WEIRD	0.74	-.12, 1.59	1.86	1.84	1.74, 1.94	0.00	0.00	100.00
Intentional Side Effects (Knobe, 2003)	1.45	.79, 2.77	1.94	1.75	1.70, 1.80	0.00	5.08	94.92
Trolley Dilemma 1 (Hauser et al., 2007)	2.50	2.22, 2.86	1.42	1.35	1.28, 1.41	0.00	0.00	100.00
False Consensus 1 (Ross et al., 1977)	0.99	0.24, 2.29	1.08	1.18	1.13, 1.23	0.00	0.00	100.00
Moral Typecasting (Gray & Wegner, 2009)	0.80	.31, 1.29	1.04	0.95	.91, 1.00	0.00	5.00	95.00
False Consensus 2 (Ross et al., 1977)	0.80	0.22, 1.87	0.89	0.95	.90, 1.00	0.00	6.67	93.33
Intuitive Reasoning (Norenzayan et al. 2002) - WEIRD	0.00	-0.15, .15	0.95	0.95	.90, 1.00	0.00	2.33	97.67
Intuitive Reasoning (Norenzayan et al. 2002) - less WEIRD	0.69	.24, 1.13	0.50	0.56	.46, .65	0.00	42.86	57.14
Less is Better (Hsee, 1998)	0.69	.24, 1.13	0.86	0.78	.74, .83	0.00	10.53	89.47
Direction & SES (Huang et al., 2014) - WEIRD	0.83	.37, 1.28	0.66	0.55	.49, .61	4.35	30.43	65.22
Direction & SES (Huang et al., 2014) - less WEIRD	-0.59	-.99, -.19	-0.10	0.03	-.05, .13	5.56	83.33	11.11
Framing (Tversky & Kahneman, 1981)	1.08	.71, 1.45	0.38	0.40	.35, .45	0.00	54.55	45.45
Moral Foundations (Graham et al., 2009)	0.52	.40, .63	0.23	0.29	.25, .34	0.00	75.00	25.00
Trolley Dilemma 2 (Hauser et al., 2007)	0.34	.26, .42	0.22	0.25	.20, .30	0.00	81.67	18.33
Tempting Fate (Risen & Gilovich, 2008)	0.39	.03, .75	0.23	0.18	.14, .22	1.69	72.88	25.42
Priming consumerism (Bauer et al., 2012)	0.87	.41, 1.34	0.16	0.12	.07, .17	1.85	87.04	11.11
Incidental Anchors (Critcher & Gilovich, 2008)	0.30	.02, .58	0.00	0.04	-.01, .09	3.39	91.53	5.08
Position & Power (Giessner & Schubert, 2007)	0.55	.05, 1.05	0.01	0.03	-.01, .08	1.69	94.92	3.39
Direction & Similarity (Tversky & Gati, 1978)	0.48	.16, .80	0.03	0.01	-.02, .04	2.04	97.96	0.00
Moral Cleansing (Zhong & Liljenquist, 2006)	1.02	.39, 2.44	0.00	0.00	-.05, .04	0.00	94.23	5.77
Structure & Goal-pursuit (Kay et al., 2014)	0.49	0.001, .973	-0.02	-0.02	-.07, .03	0.00	100.00	0.00
Social Value Orientation (Van Lange et al., 1997)	0.19	<.001, .47	0.06	-0.03	-.08, .02	0.00	98.15	1.85
Priming warmth affects climate beliefs (Zaval et al., 2014)	0.31	.03, .59	0.00	-0.03	-.09, .03	5.36	89.29	5.36
Incidental Disfluency (Alter et al., 2007)	0.63	-.004, 1.25	-0.07	-0.03	-.08, .01	1.52	96.97	1.52
SMS & Well-Being (Anderson et al., 2012)	0.57	.20, .93	-0.05	-0.04	-.09, -.004	0.00	94.92	5.08
Choosing or Rejecting (Shafir, 1993)	0.35	-.04, .68	-0.04	-0.13	-.18, -.09	18.97	79.31	1.72
Affect & Risk (Rottenstreich & Hsee, 2001)	0.74	<.001, 1.74	-0.06	-0.08	-.13, -.03	3.33	95.00	1.67
Actions are Choices (Savani et al. 2010) - WEIRD	0.08	-.33, .50	-0.24	-0.21	-.23, -.18	46.51	53.49	0.00
Actions are Choices (Savani et al. 2010) - less WEIRD	-0.65	-1.01, -.30	-0.14	-0.12	-.16, -.08	28.57	71.43	0.00

△ **Original Effect Size**

Cohen's q

−3 −2 −1 0 1 2 3

Disgust & Homophobia (Inbar et al., 2009)

Assimilation & Contrast (Schwarz et al., 1991)

Correspondence Bias (Miyamoto & Kitayama, 2002)

Intentional Side−Effects (Knobe, 2003)

Trolley Dilemma 1 (Hauser et al., 2007)

False Consensus 1 (Ross et al., 1977)

Moral Typecasting (Gray & Wegner, 2009)

False Consensus 2 (Ross et al., 1977)

Intuitive Reasoning (Norenzayan et al., 2002)

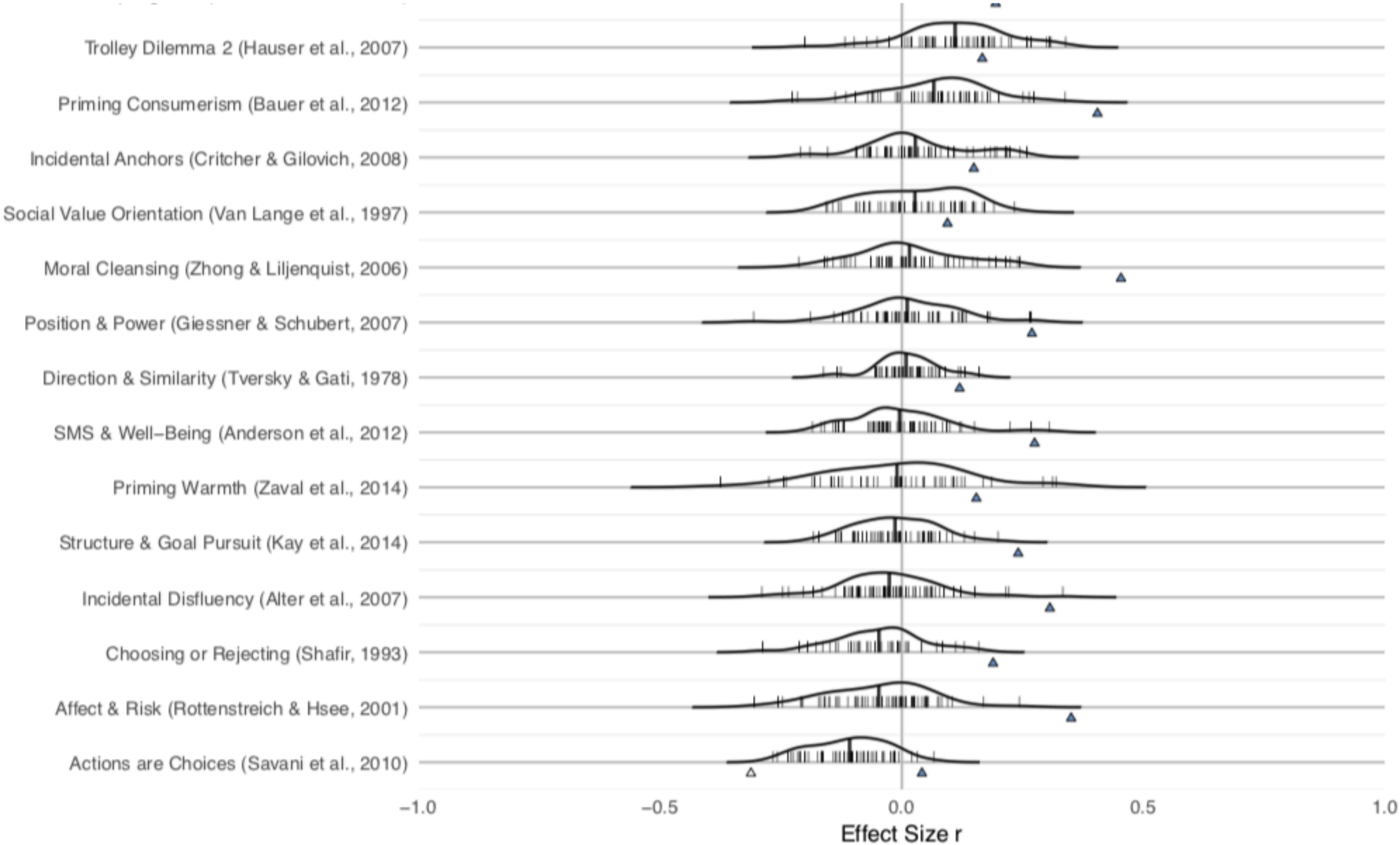
Less is Better (Hsee, 1998)

Framing (Tversky & Kahneman, 1981)

Direction & SES (Huang et al., 2014)

Moral Foundations (Graham et al., 2009)

Tempting Fate (Risen & Gilovich, 2008)


Trolley Dilemma 2 (Hauser et al., 2007)
Priming Consumerism (Bauer et al., 2012)
Incidental Anchors (Critcher & Gilovich, 2008)
Social Value Orientation (Van Lange et al., 1997)
Moral Cleansing (Zhong & Liljenquist, 2006)
Position & Power (Giessner & Schubert, 2007)
Direction & Similarity (Tversky & Gati, 1978)
SMS & Well–Being (Anderson et al., 2012)
Priming Warmth (Zaval et al., 2014)
Structure & Goal Pursuit (Kay et al., 2014)
Incidental Disfluency (Alter et al., 2007)
Choosing or Rejecting (Shafir, 1993)
Affect & Risk (Rottenstreich & Hsee, 2001)
Actions are Choices (Savani et al., 2010)
−1.0
−0.5
0.0
0.5
1.0
Effect Size r

What’s gone wrong?

The Academic Incentive Structure

We live in a publish or perish culture.

Publication number, where you publish, and citations are all used (either explicitly or implicitly) in appointment and promotion committees.

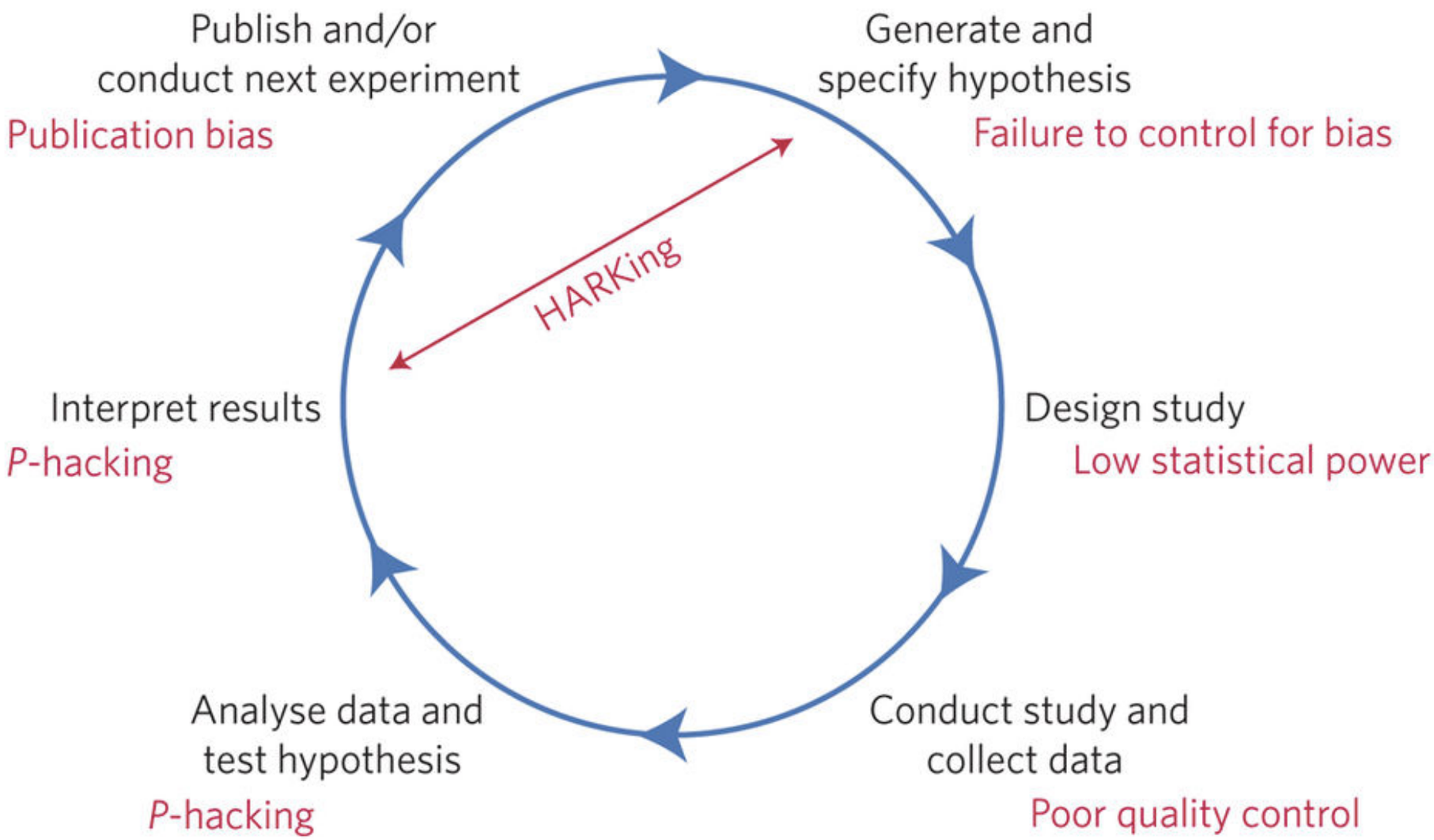
REF's definition of 3* and 4* research (although this looks like it could be changing).

Is there not just “good science” and “bad science”?

Without realising it, good scientists have been engaging in questionable research practices (QRPs) partly driven by an incentive structure that doesn’t incentivise good scientific practice…

Problems include *p*-hacking, lack of power, HARKing, failing (refusal) to share data and code, too many researcher degrees of freedom…

From: A manifesto for reproducible science



Munafo et al. (2017), *Nature Human Behaviour*

It’s not a new problem…

“*I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories . . . is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology.*” Paul Meehl, 1978.

https://osf.io/teba2/

ANNUAL REVIEWS

Annual Review of Psychology

Psychology's Renaissance

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

[1]Haas School of Business, University of California, Berkeley, California 94720; email: Leif_Nelson@haas.berkeley.edu

[2]The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; email: jsimmo@upenn.edu, urisohn@gmail.com

"the overwhelming majority of published findings are statistically significant (Fanelli 2012, Greenwald 1975, Sterling 1959). On the other hand, the overwhelming majority of published studies are underpowered and, thus, theoretically unlikely to obtain results that are statistically significant."

Personality and Social Psychology Review
1998, Vol. 2, No. 3, 196–217

HARKing: Hypothesizing After the Results are Known

Norbert L. Kerr
Department of Psychology
Michigan State University

This article considers a practice in scientific communication termed HARKing (Hypothesizing After the Results are Known). HARKing is defined as presenting a post hoc hypothesis (i.e., one based on or informed by one's results) in one's research report as if it were, in fact, an a priori hypotheses. Several forms of HARKing are identified and survey data are presented that suggests that at least some forms of HARKing are widely practiced and widely seen as inappropriate. I identify several reasons why scientists might HARK. Then I discuss several reasons why scientists ought not to HARK. It is conceded that the question of whether HARKing's costs exceed its benefits is a complex one that ought to be addressed through research, open discussion, and debate. To help stimulate such discussion (and for those such as myself who suspect that HARKing's costs do exceed its benefits), I conclude the article with some suggestions for deterring HARKing.

ROBERT TAYLOR

Rein in the four horsemen of irreproducibility

Dorothy Bishop *describes how threats to reproducibility, recognized but unaddressed for decades, might finally be brought under control.*

More than four decades into my scientific career, I find myself an outlier among academics of similar age and seniority: I strongly identify with the movement to make the practice of science more robust. It's not that my contemporaries are unconcerned about doing science well; it's just that many of them don't seem to recognize that there are serious problems with current practices. By contrast, I think that, in two decades, we will look back on the past 60 years — particularly in biomedical science — and marvel at how much time and money has been wasted on flawed research.

How can that be? We know how to formulate and test hypotheses in controlled experiments. We can account for unwanted variation with statistical techniques. We appreciate the need to replicate observations.

Yet many researchers persist in working in a way almost guaranteed not to deliver meaningful results. They ride with what I refer to as the four horsemen of the reproducibility apocalypse: publication bias, low statistical power, *P*-value hacking and HARKing (hypothesizing after results are known). My generation and the one before us have done little to rein these in.

In 1975, psychologist Anthony Greenwald noted that science is prejudiced against null hypotheses; we even refer to sound work supporting such conclusions as 'failed experiments'. This prejudice leads to publication bias: researchers are less likely to write up studies that show no effect, and journal editors are less likely to accept them. Consequently, no one can learn from them, and researchers waste time and resources be adequately powered. Other disciplines have yet to catch up.

I stumbled on the issue of *P*-hacking before the term existed. In the 1980s, I reviewed the literature on brain lateralization (how sides of the brain take on different functions) and developmental disorders, and I noticed that, although many studies described links between handedness and dyslexia, the definition of 'atypical handedness' changed from study to study — even within the same research group. I published a sarcastic note, including a simulation to show how easy it was to find an effect if you explored the data after collecting results (D. V. M. Bishop *J. Clin. Exp. Neuropsychol.* **12**, 812–816; 1990). I subsequently noticed similar phenomena in other fields: researchers try out many analyses but report only the ones that are 'statistically significant'.

This practice, now known as *P*-hacking, was once endemic to most branches of science that rely on *P* values to test significance of results, yet few people realized how seriously it could distort findings. That started to change in 2011, with an elegant, comic paper in which the authors crafted analyses to prove that listening to the Beatles could make undergraduates younger (J. P. Simmons *et al. Psychol. Sci.* **22**, 1359–1366; 2011). "Undisclosed flexibility," they wrote, "allows presenting anything as significant."

The term HARKing was coined in 1998 (N. L. Kerr *Pers. Soc. Psychol. Rev.* **2**, 196–217; 1998). Like *P*-hacking, it is so widespread that researchers assume it is good practice. They look at the data, pluck out a finding that looks exciting and write a paper to tell a story around this result. Of course, researchers should be free to explore their

MANY RESEARCHERS PERSIST IN WORKING IN A WAY ALMOST GUARANTEED NOT TO DELIVER MEANINGFUL RESULTS.

https://www.nature.com/articles/d41586-019-01307-2

Distinguishing between replicability and reproducibility (note, both are important!)

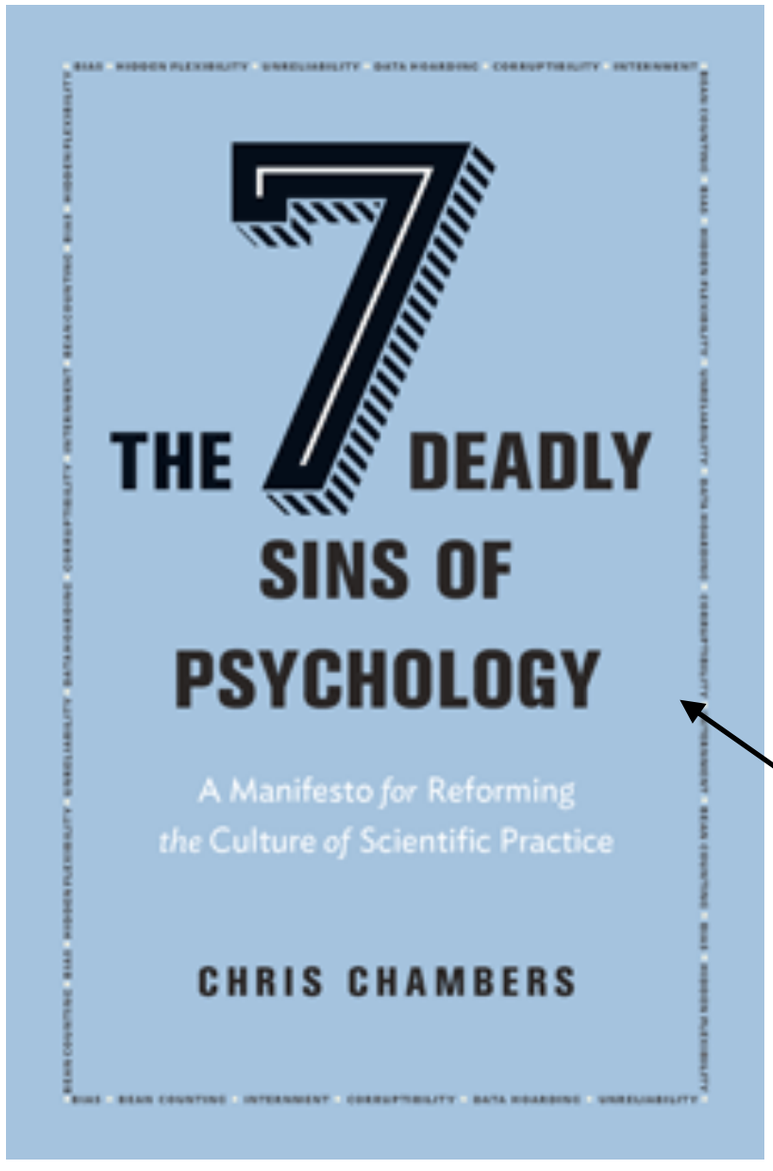
Replicable Science is when someone else can run a study the same as or conceptually equivalent to your one, and find a similar pattern of effects.

Reproducible Science is when someone else can take your data and your analysis code, run it and then find the same effects that you have reported.

How do we make our science more replicable?

How do we make our science more reproducible?

A move towards open science…



Sins include *p*-hacking, lack of power, HARKing, failing (refusal) to share data and code, too many researcher degrees of freedom…

You really should read this book!

http://www.stat.columbia.edu/~gelman/

Andrew Gelman gives the following recommendations to researchers:

- Analyze all your data.
- Present all your comparisons.
- Make your data public.
- Put in the effort to take accurate measurements (low bias, low variance, and a large enough sample size).
- Do repeated-measures comparisons where possible.

Open Science practices include…

- Pre-registering experiments.
- Registered reports.
- Using preprint servers (e.g., bioRxiv, PsyArXiv).
- Making data and analysis code freely available (e.g., via GitHub, OSF).
- Open access to journal articles.
- …and more.

Open Science recently recognised by G7 Science Ministers…

Focus: Incentives and the researcher ecosystem

Ambition: Foster a research environment in which career advancement takes into account Open Science activities, through incentives and rewards for researchers, and valuing the skills and capabilities in the Open Science workforce.

Recommendations:

At national levels: G7 nations should each engage with research stakeholders to identify and implement enhancements to research evaluation and reward systems that take into consideration the Open Science activities carried out by researchers and research institutions. Topics that could be discussed include:

- Recognizing Open Science practices during evaluation of research funding proposals, and research outcomes;
- Recognizing and rewarding research productivity and impact that reflect open science activities by researchers during career advancement reviews;
- Including credit for service activities such as reviewing, evaluating, and curation and management of research data; and,
- Developing metrics of Open Science practices.

In REF2021 UoA Environment…

29. The revised template will also include a **section on 'open research'**, detailing the submitting unit's open access strategy, including where this goes above and beyond the REF open access policy requirements, and wider activity to encourage the effective sharing and management of research data. The panels will set out further guidance on this in the panel criteria.

is beginning to appear in tenure-track job adverts…

> Our Department embraces the values of open and reproducible science, and candidates are encouraged to address (in their statements and/or cover letter) how they have pursued and/or plan to pursue these goals in their work.

and is forming part of Universities' teaching manifestos.

Teaching with Open Science commitment:
To teach the practices and skills of open research and science in our undergraduate and postgraduate degree programmes

a. Promote open science in our teaching.
b. Design a Research Methods curriculum that teaches skills for open science and uses open science to enhance teaching (for example: teach R and use open data to practice analysis skills).
c. Learn about and adopt open educational practices in our teaching.
d. Produce and promote tools for helping student researchers adopt open practices, including training and guidance suitable to their level of study.
e. Author, share and use open educational resources to promote teaching with open science beyond our School and Institution.
f. Support our colleagues to learn the skills of teaching Open Science.

Part of doing better science involves knowing how to build appropriate statistical models, and how to understand what those models are telling you (and what they are not…)

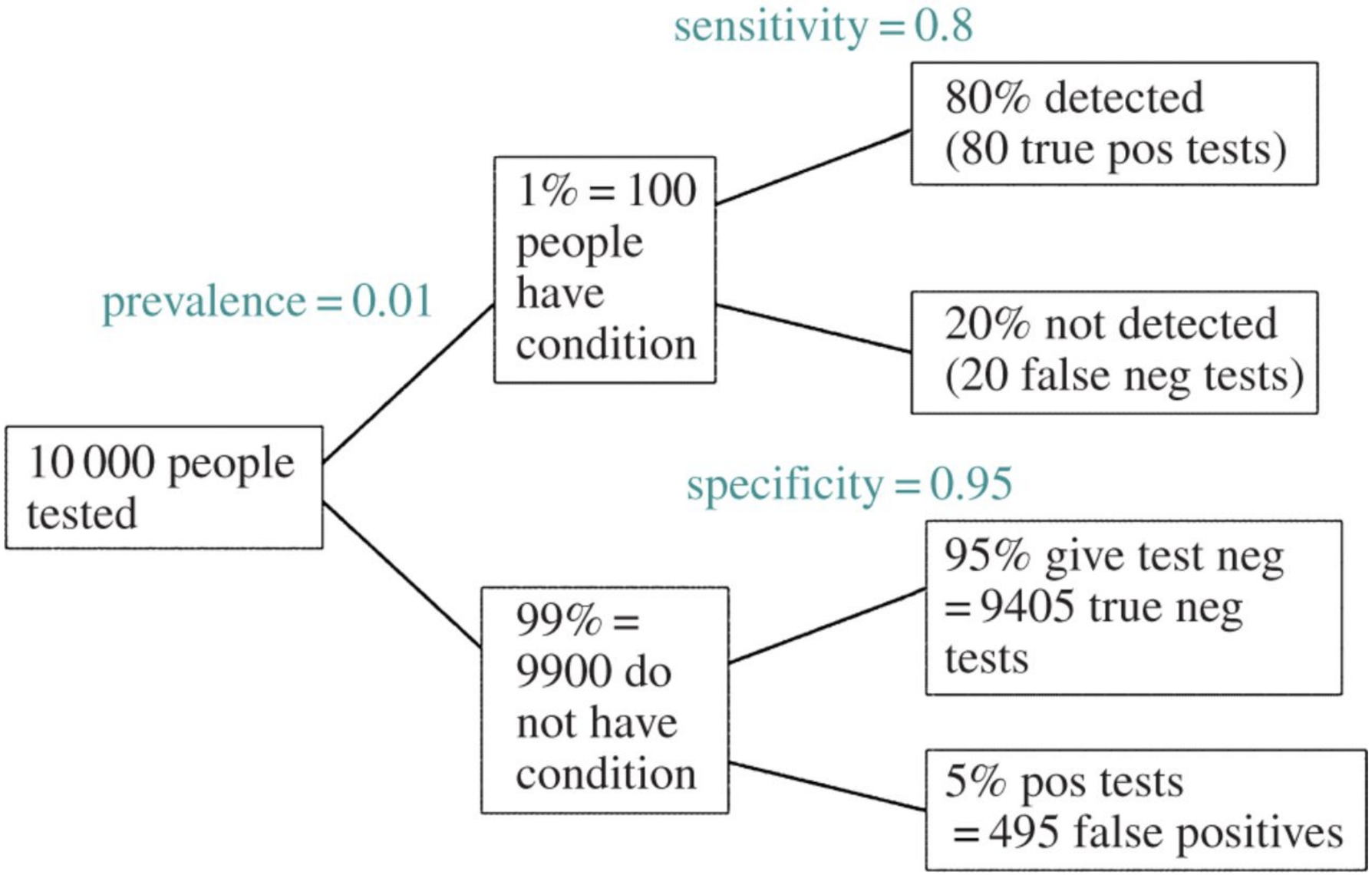
Why Understanding Statistics Matters…

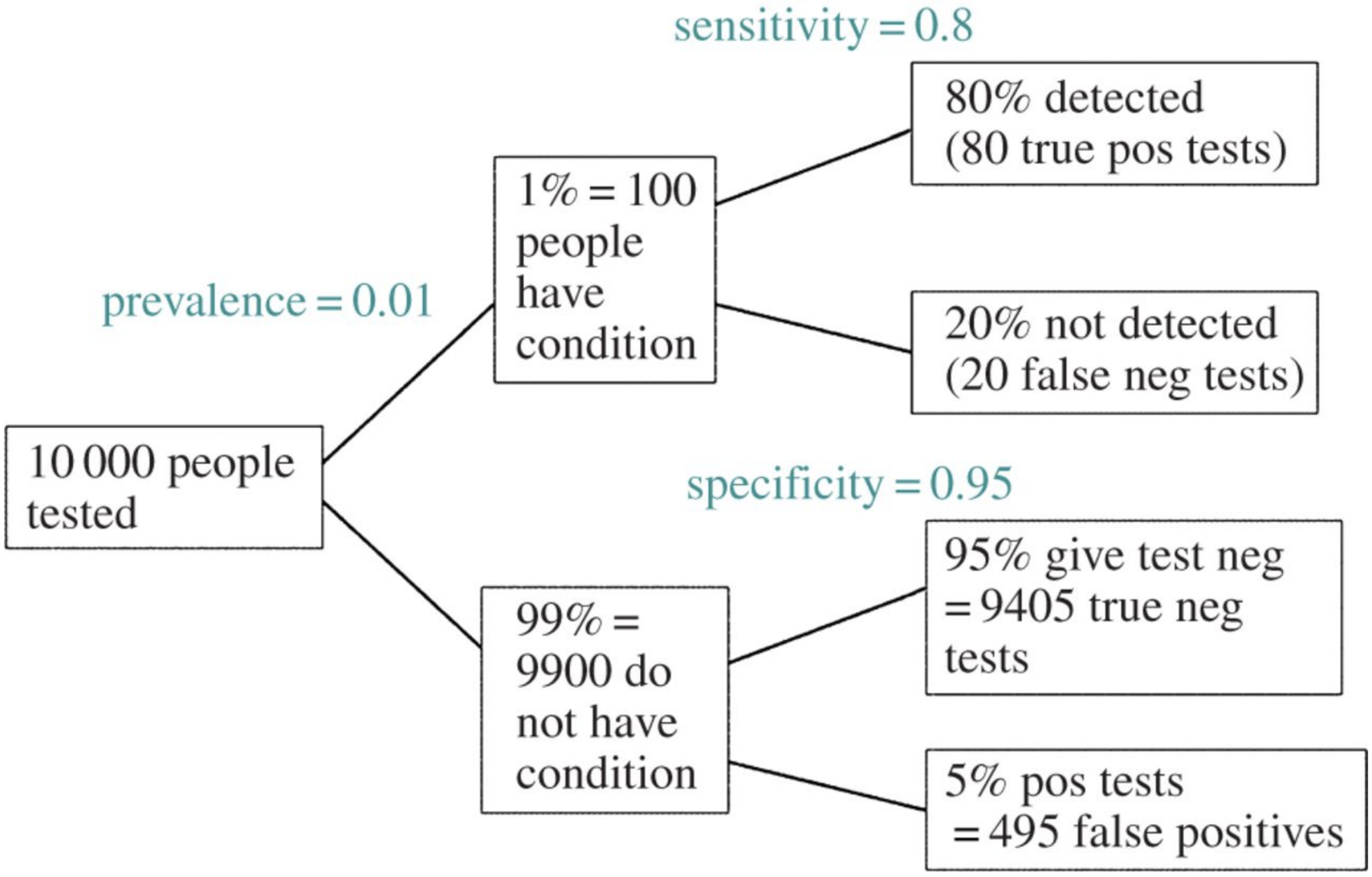
- Imagine a test in which 95% of people without a medical condition will be correctly diagnosed as not having it (specificity = 0.95).

- Imagine the test is able to correctly diagnose 4 out of the 5 people who **do** have the medical condition (sensitivity = 0.8).

- Imagine the prevalence of the medical condition in the population is 1%.

From Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. DOI: 10.1098/rsos.140216


sensitivity = 0.8
80% detected
(80 true pos tests)
1% = 100
people
have
condition
prevalence = 0.01
20% not detected
(20 false neg tests)
10 000 people
tested
specificity = 0.95
95% give test neg
= 9405 true neg
tests
99% =
9900 do
not have
condition
5% pos tests
= 495 false positives



- The results of the test suggest 575 people have the condition. But 495 of these are false positives! So 86% of the people who produced a positive result actually don't have the condition.

Understanding Statistics

- Appropriately powered studies for the effect size of interest, appropriately analysed.
- Consider data simulation prior to data collection (does my design provide me with the richness I need to build my model and detect the minimal effect size of interest?)
- Consider additions and alternatives to NHST where appropriate.
- Recognition that our research should focus on revealing *what effects are likely to be real*, rather than just statistical significance. We need to remember what significance is (and what it isn't).

ASA Principles on *p*-values

1. *p*-values can indicate how incompatible the data are with a specified statistical model.

2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

4. Proper inference requires full reporting and transparency.

5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

"*All models are wrong,*
but some are useful”,

George Box

How do you do Open Science?

Before Data Collection

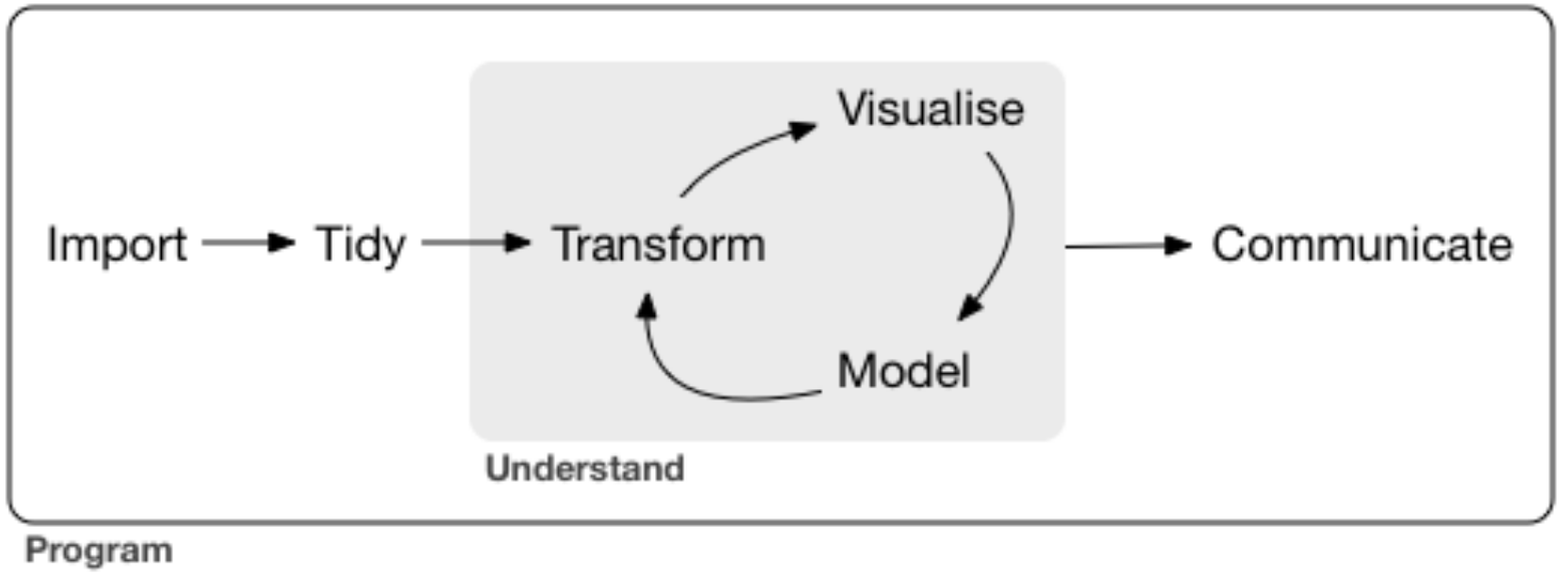
- Specify your hypotheses and analysis plan.
- **Pre-register** your hypotheses and analysis plan at `osf.io`
- Consider data simulation so that you can write your analysis script before you have your real data.
- Consider submitting as a **registered report** - currently **186** journals now support this route. This involves acceptance in principle before you have even started collecting your data.

Registered Reports

https://cos.io/rr/?_ga=2.49773158.1336120275.1555407527-1361001319.1494339346

After Data Collection

- You need to use analysis software that allows for open sharing and reproducibility of the entire data wrangling/ analysis/write-up workflow.



Hadley Wickham and Garrett Grolemund

- You can share your data at `osf.io` or on GitHub:

ajstewartlang / **Comprehension-of-indirect-requests-is-influenced-by-their-degree-of-imposition**

Watch 0 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Branch: master ▾ **Comprehension-of-indirect-requests-is-influenced-by-their-degree-of-imposition** / **RPs.csv**

Find file | Copy path

ajstewartlang Made consistent the labelling of factors in data files and in paper — 7b3b3b1 on 29 Mar 2017

0 contributors

1681 lines (1681 sloc) | 69.7 KB — Raw | Blame | History

Search this file...

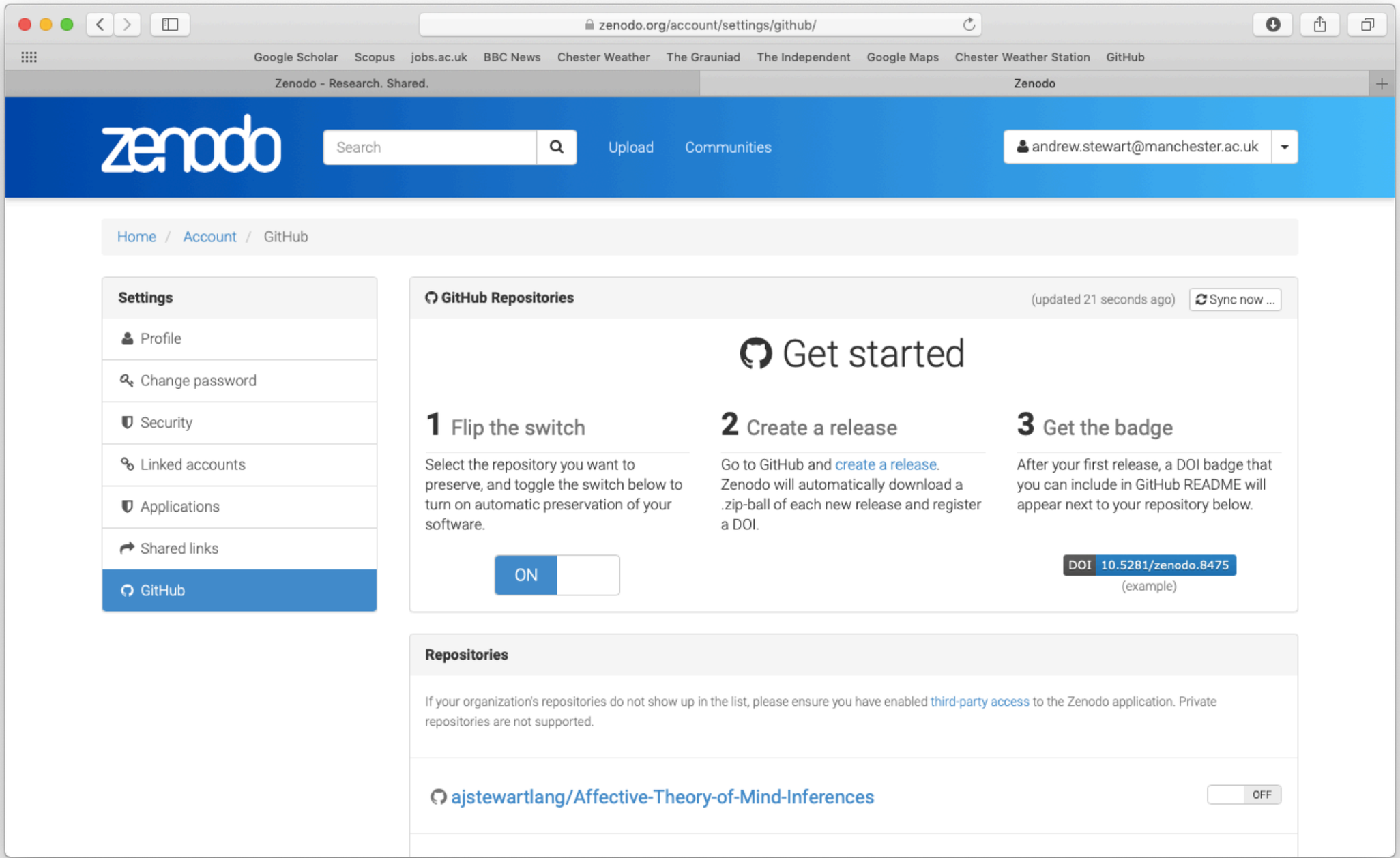
	P.s	Item	Condition	Probmanip	Speaker	statement	response	final	Meaning	Imposition
2	1	1	1	1708	302	1399	1867	1206	Indirect	High
3	1	2	2	1466	296	1377	1674	828	Indirect	Low
4	1	3	3	1393		1494	1950	1812	Direct	High
5	1	4	4	2463	530	1691	1866	965	Direct	Low
6	1	5	1	1552	267	1332	1477	1345	Indirect	High
7	1	6	2	1445	444	1004	1067	797	Indirect	Low
8	1	7	3	2159	501	739	1231	2240	Direct	High
9	1	8	4	1459		1086	946	978	Direct	Low
10	1	9	1	3302		1503	900	1736	Indirect	High

FaceTime

- alongside your analysis code

```
FPs$Meaning <- as.factor(FPs$Meaning)
FPs$Imposition <- as.factor(FPs$Imposition)

#this sets up the contrasts so that the intercept in the mixed LMM is the grand mean (i.e., the mean of all conditions)
my.coding <- matrix (c(.5, -.5))

contrasts (FPs$Meaning) <- my.coding
contrasts (FPs$Imposition) <- my.coding

#construct the models with crossed random effects for subjects and items for the pre-critical, critical and post-crtical region
fpmodelprec <- lmer (Probmanip ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning+Imposition |Item), data=FPs, REM
summary (fpmodelprec)
lsmeans (fpmodelprec, pairwise~Meaning*Imposition, adjust="none")

fpmodelc <- lmer (statement ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning*Imposition |Item), data=FPs, REML=T
summary (fpmodelc)
lsmeans (fpmodelc, pairwise~Meaning*Imposition, adjust="none")

fpmodelpostc <- lmer (response ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning+Imposition |Item), data=FPs, REM
summary (fpmodelpostc)
lsmeans (fpmodelpostc, pairwise~Meaning*Imposition, adjust="none")

#Regression Path Analysis
#Read in Regression Path data
RPs <- read.csv("~/RPs.csv")

RPs$Meaning <- as.factor(RPs$Meaning)
RPs$Imposition <- as.factor(RPs$Imposition)

contrasts (RPs$Meaning) <- my.coding
contrasts (RPs$Imposition) <- my.coding

#construct the models with crossed random effects for subjects and items for the pre-critical, critical and post-crtical region
rpmodelprec <- lmer (Probmanip ~ Meaning*Imposition + (1+Meaning*Imposition |P.s) + (1+Meaning*Imposition |Item), data=RPs, REM
```

And preserve it with a DOI via Zenodo



Using R for Data Analysis

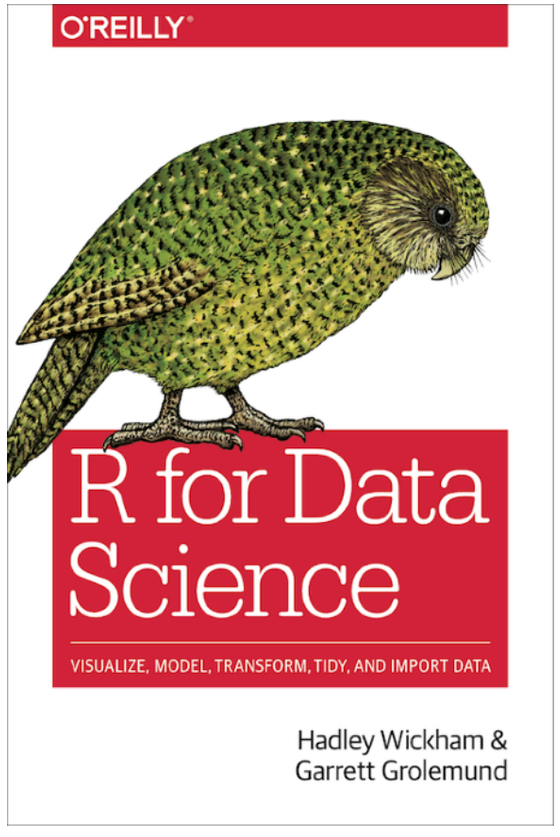
“Hadley Wickham, the Man Who Revolutionized R”

Chief Scientist at RStudio, author of key R packages incl. `ggplot2`, `tidyr`, `dplyr` - all components of the tidyverse.

https://priceonomics.com/hadley-wickham-the-man-who-revolutionized-r/

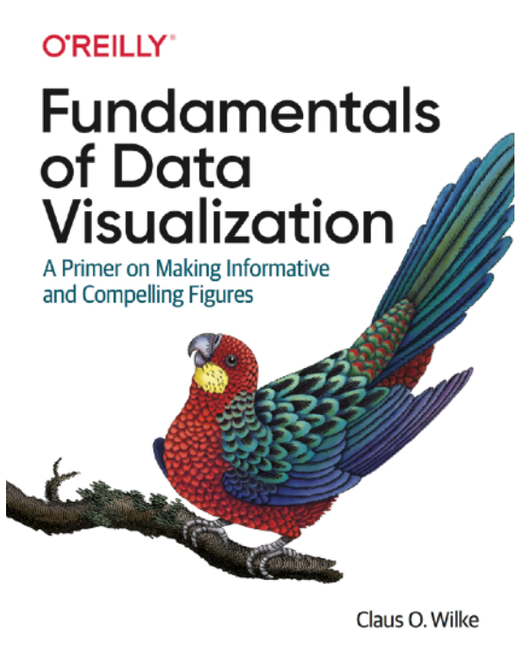
What role can R play in Open Science?

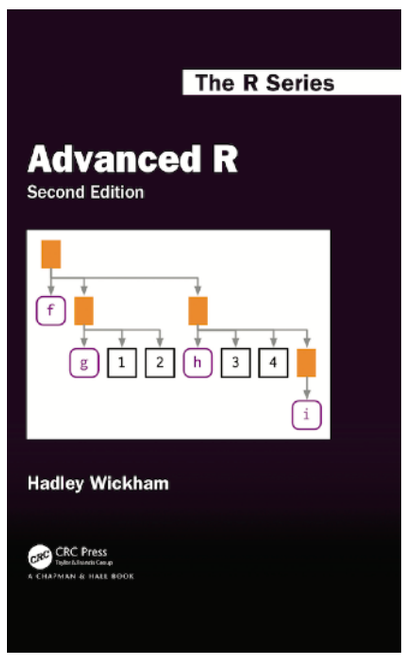
- R scripts are easy to share allowing for reproducibility and easy public sharing of data and code.
- R is free, open source software that is much more flexible and powerful than SPSS.
- There is an active R community continuously updating statistical tests and packages that run in R.
- As R is a programming language, it forces you to know your data.

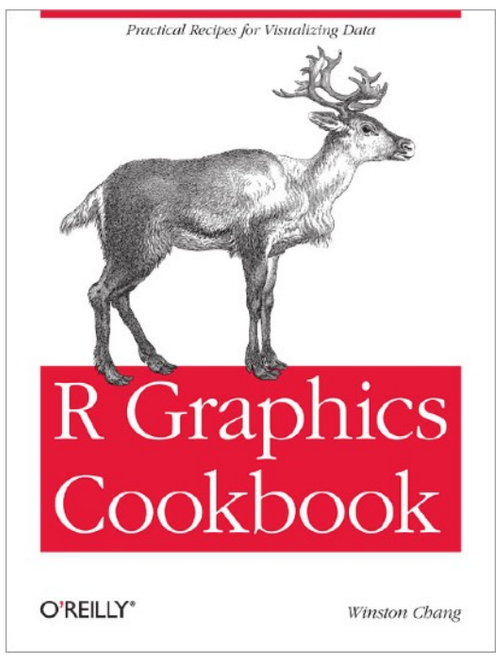


Available electronically for free at:

`http://r4ds.had.co.nz`







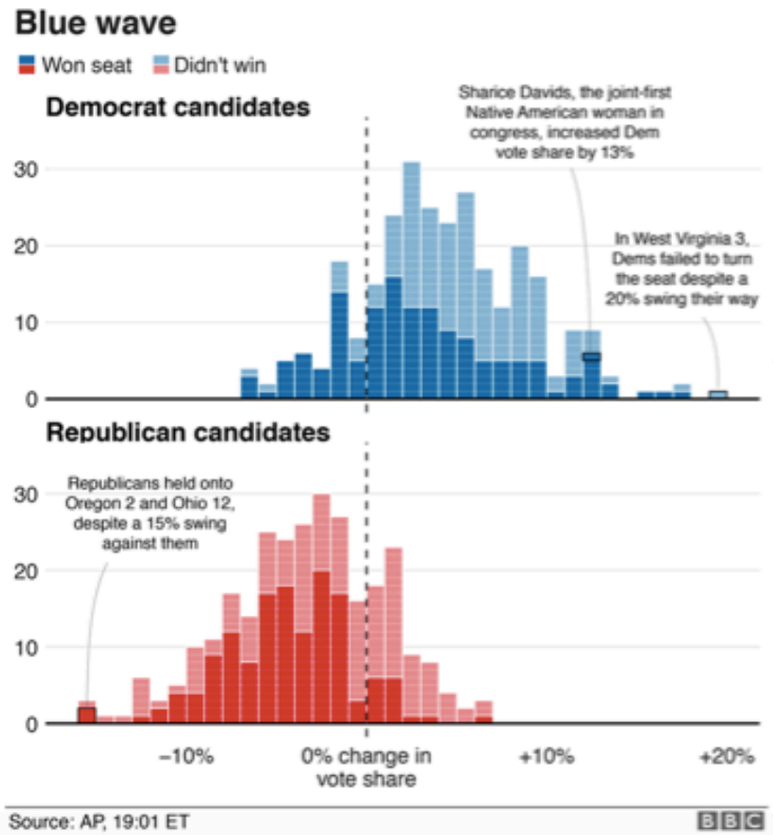
BBC Visual and Data Journalism cookbook for R graphics

Last updated: 2019-01-24

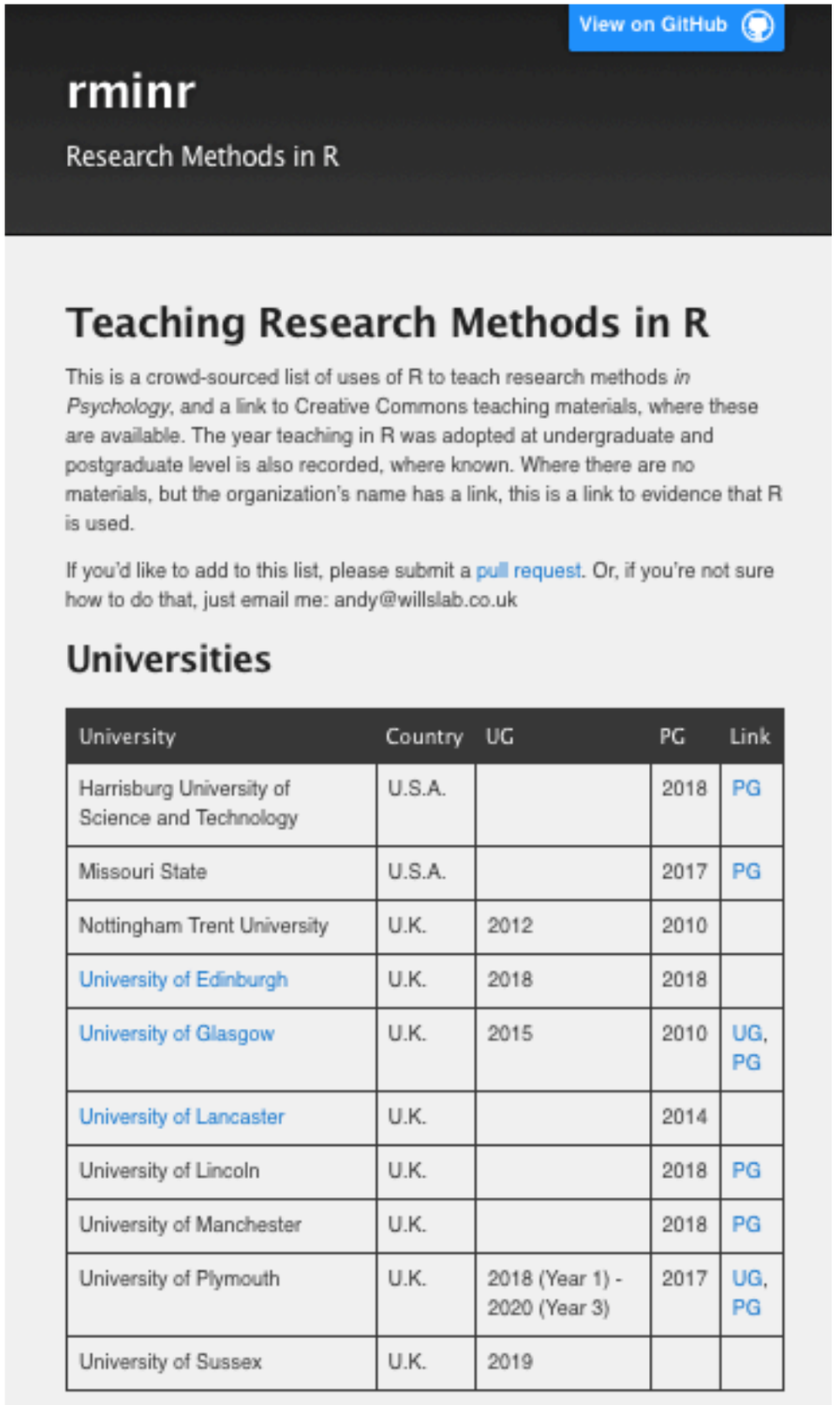
How to create BBC style graphics

At the BBC data team, we have developed an R package and an R cookbook to make the process of creating publication-ready graphics in our in-house style using R's ggplot2 library a more reproducible process, as well as making it easier for people new to R to create graphics.

The cookbook below should hopefully help anyone who wants to make graphics like these:



Handy list of Psychology groups that teach R, plus links to course materials - list compiled by Andy Wills at Plymouth.



View on GitHub

rminr

Research Methods in R

Teaching Research Methods in R

This is a crowd-sourced list of uses of R to teach research methods *in Psychology*, and a link to Creative Commons teaching materials, where these are available. The year teaching in R was adopted at undergraduate and postgraduate level is also recorded, where known. Where there are no materials, but the organization's name has a link, this is a link to evidence that R is used.

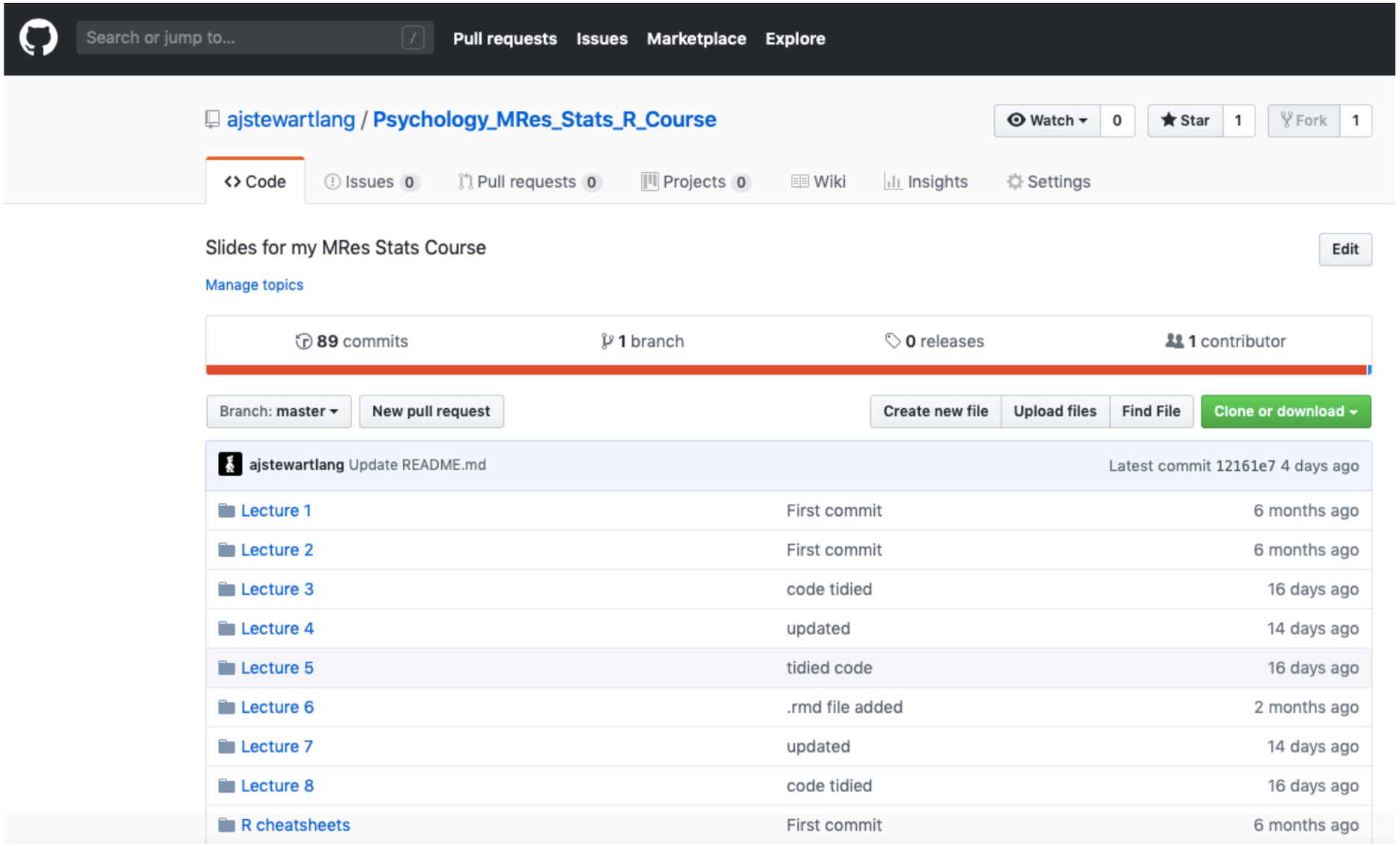
If you'd like to add to this list, please submit a pull request. Or, if you're not sure how to do that, just email me: andy@willslab.co.uk

Universities

University	Country	UG	PG	Link
Harrisburg University of Science and Technology	U.S.A.		2018	PG
Missouri State	U.S.A.		2017	PG
Nottingham Trent University	U.K.	2012	2010	
University of Edinburgh	U.K.	2018	2018	
University of Glasgow	U.K.	2015	2010	UG, PG
University of Lancaster	U.K.		2014	
University of Lincoln	U.K.		2018	PG
University of Manchester	U.K.		2018	PG
University of Plymouth	U.K.	2018 (Year 1) - 2020 (Year 3)	2017	UG, PG
University of Sussex	U.K.	2019		

https://ajwills72.github.io/rminr/rminrinpsy.html

My (free!) M-Level R Course For Psychologists



https://github.com/ajstewartlang/Psychology_MRes_Stats_R_Course

Journals recognise OS practices

VIEW THE **BADGES:**

Brian Nosek
@BrianNosek

Following

The Power of Norms: Every single article in this month's Psychological Science earned an open data badge.

8/14 open materials badge, and 5/14 preregistration badge.

Four triple badgers in a single issue.

<swoon>

SCIENCE

A JOURNAL OF THE ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

OPEN DATA · OPEN MATERIALS · PREREGISTERED

Volume 30 | Number 4 | April 2019

The links below take you to the journal via the APS website. If not already logged in, you will be redirected to log-in using your last name and Member ID (16345). Be sure to download pictures to see Open Science badges.

Unconscious Detection of One's Own Image
Michał J. Wójcik, Maria M. Nowicka, Michał Bola, and Anna Nowicka

Using a Psychopharmacogenetic Approach To Identify the Pathways Through Which—and the People for Whom—Testosterone Promotes Aggression
Shawn N. Geniole, Tanya L. Procyshyn, Nicole Marley, Triana L. Ortiz, Brian M. Bird, Ashley L. Marcellus, Keith M. Welker, Pierre L. Bonin, Bernard Goldfarb, Neil V. Watson, and Justin M. Carré

The Additive-Area Heuristic: An Efficient but Illusory Means of Visual Area Approximation
Sami R. Yousif and Frank C. Keil

Preschoolers Optimize the Timing of Their Conversational Turns Through Flexible Coordination of Language Comprehension and Production
Laura Lindsay, Chiara Gambi, and Hugh Rabagliati

Causal Inference About Good and Bad Outcomes
Hayley M. Dorfman, Rahul Bhui, Brent L. Hughes, and Samuel J. Gershman

Dissecting the Neural Focus of Attention Reveals Distinct Processes for Spatial Attention

Racial Bias in Perceptions of Size and Strength: The Impact of Stereotypes and Group Differences
David J. Johnson and John Paul Wilson

Property Damage and Exposure to Other People in Distress Differentially Predict Prosoc Behavior After a Natural Disaster
Tom Vardy and Quentin D. Atkinson

The Prevalence of Marginally Significant Results in Psychology Over Time
Anton Olsson-Collentine, Marcel A. L. M. van Assen, and Chris H. J. Hartgerink

Reactivation of Previous Experiences in a Working Memory Task
Gi-Yeul Bae and Steven J. Luck

Group-Based Relative Deprivation Explains Endorsement of Extremism Among Western-Born Muslims
Milan Obaidi, Robin Bergh, Nazar Akrami, and Gulnaz Anjum

Null Effects of Game Violence, Game Difficulty, and 2D:4D Digit Ratio on Aggressive Behavior
Joseph Hilgard, Christopher R. Engelhardt, Jeffrey N. Rouder, Ines L. Segert, and Bruce D. Bartholow

Collective Emotions and Social Resilience in the Digital Traces After a Terrorist Attack
David Garcia and Bernard Rimé

3:13 PM - 17 Apr 2019

Registered reports are fundamentally changing the shape of the publishing landscape.

NEWS · 24 OCTOBER 2018

First analysis of 'pre-registered' studies shows sharp rise in null findings

Logging hypotheses and protocols before performing research seems to work as intended: to reduce publication bias for positive results.

https://www.nature.com/articles/d41586-018-07118-1

Other considerations if you want to do more open and reproducible science…

Sharing your computational environment

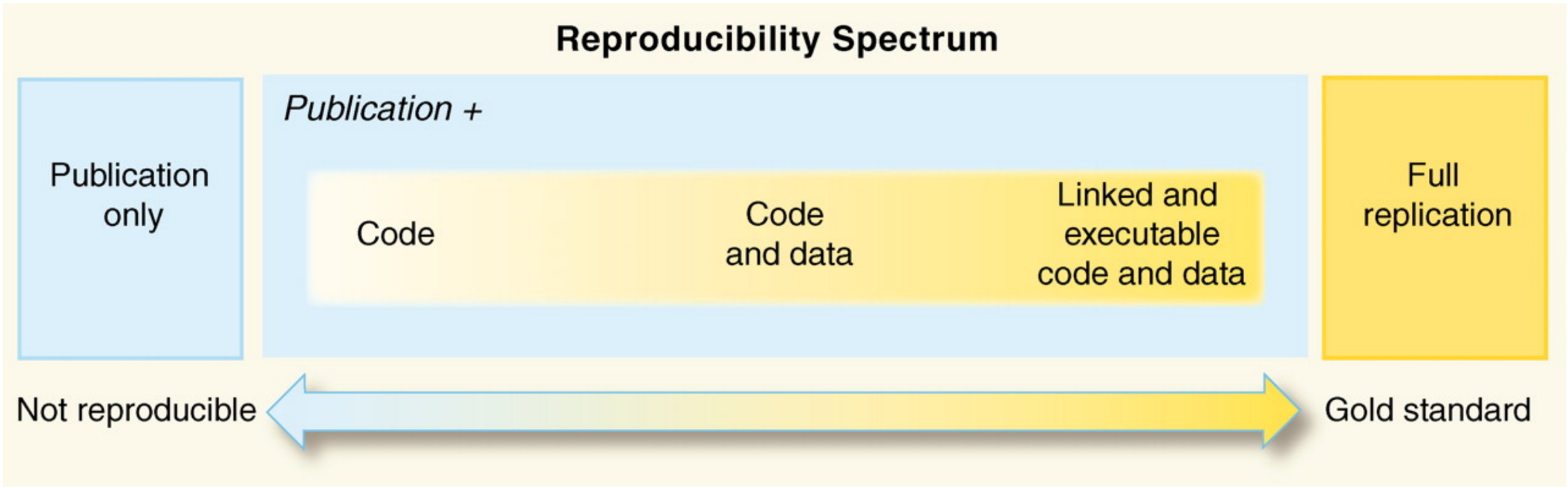
PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

+ See all authors and affiliations

Science 02 Dec 2011:
Vol. 334, Issue 6060, pp. 1226-1227
DOI: 10.1126/science.1213847



Consider a multiverse analytical approach

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2016, Vol. 11(5) 702–712

DOI: 10.1177/1745691616658637
pps.sagepub.com
SAGE

Increasing Transparency Through a Multiverse Analysis

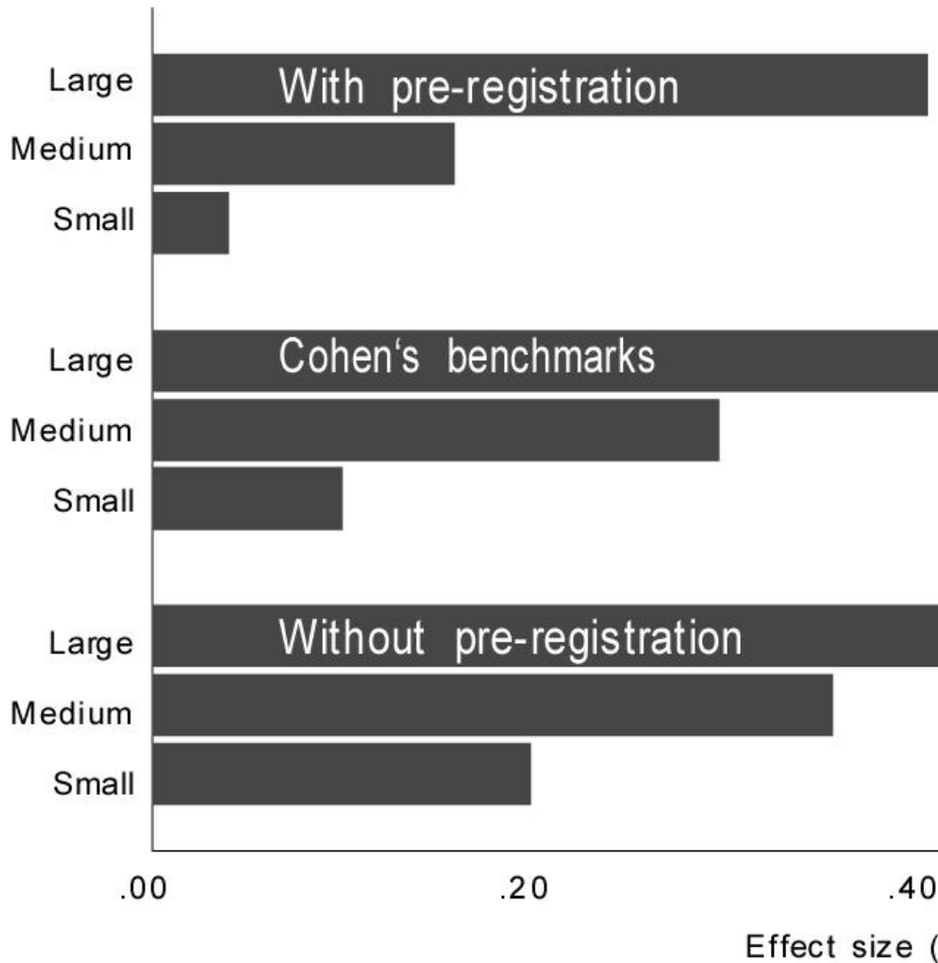
Sara Steegen[1], Francis Tuerlinckx[1], Andrew Gelman[2], and Wolf Vanpaemel[1]

[1]KU Leuven, University of Leuven and [2]Columbia University

Abstract

Empirical research inevitably includes constructing a data set by processing raw data into a form ready for statistical analysis. Data processing often involves choices among several reasonable options for excluding, transforming, and coding data. We suggest that instead of performing only one analysis, researchers could perform a multiverse analysis, which involves performing all analyses across the whole set of alternatively processed data sets corresponding to a large set of reasonable scenarios. Using an example focusing on the effect of fertility on religiosity and political attitudes, we show that analyzing a single data set can be misleading and propose a multiverse analysis as an alternative practice. A multiverse analysis offers an idea of how much the conclusions change because of arbitrary choices in data construction and gives pointers as to which choices are most consequential in the fragility of the result.

Realise that actual effect sizes may be much smaller than Cohen thought...



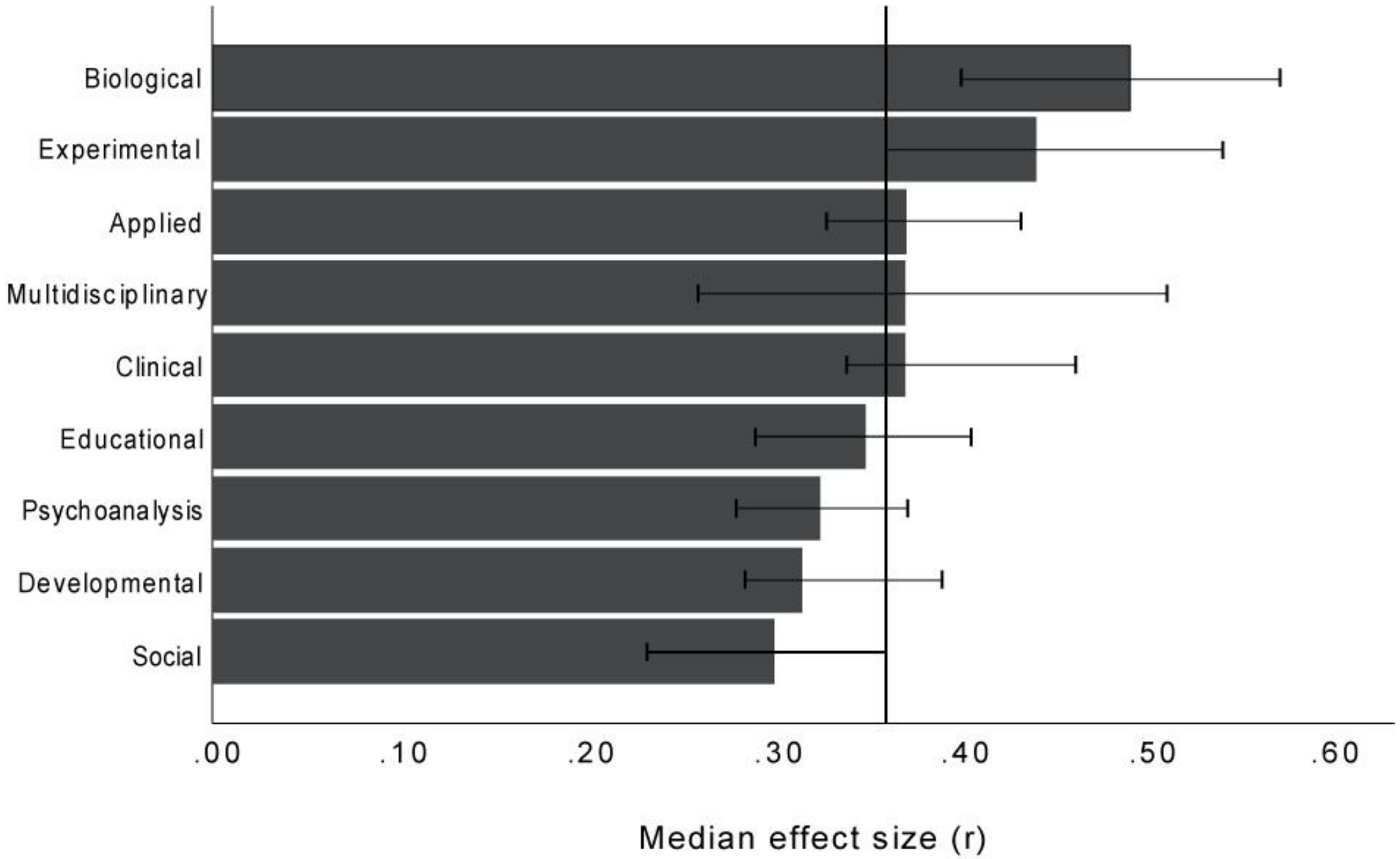
ORIGINAL RESEARCH ARTICLE
Front. Psychol., 11 April 2019 | https://doi.org/10.3389/fpsyg.2019.00813

The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases

Thomas Schäfer* and Marcus A. Schwarz

Department of Psychology, Chemnitz University of Technology, Chemnitz, Germany

With lots of variability between sub-disciplines…



Set up your own Open Science Working Group

- Open Science Working Group at Manchester founded in November by myself and Caroline Jay (Computer Science) - subscribe to our listserve:

https://listserv.manchester.ac.uk/cgi-bin/wa?REPORT=OPEN_RESEARCH

- Lots of OS activities incl. reproducibility journal club (ReproducibiliTea) meeting fortnightly.

- Visit from Dorothy Bishop next year (Feb 26th) to talk about reproducibility - you're all invited!

- Check out the Network of Open Science Working groups: https://osf.io/vgt3x/

North West Open Science Network

- We are part of a broader network in the NW including Lancaster, Keele, MMU.

- We are also part of the UK Reproducibility Network funded/supported by UKRI, research England, MRC, NERC, ESRC, Wellcome, Universities UK, JISC, British Neuroscience Association (amongst others).

- Links to Project Tier, The Carpentries, Software Sustainability Institute, The Turing Way etc.

The UK Reproducibility Network

The power of networks

A group of researchers recently launched the UK Reproducibility Network, supported by Jisc and a range of other stakeholders, including funders and publishers.

Our aim is to bring together colleagues across the higher education and research sector, forming local networks at individual institutions to promote the adoption of initiatives intended to improve research.

> This is very much a peer-led, grassroots initiative that will allow academics to coordinate their efforts and engage with key stakeholders.

https://www.jisc.ac.uk/blog/how-do-you-deal-with-a-problem-like-reproducibility-29-nov-2018

Project TIER

The Journal of Economic Education
Journal
The Journal of Economic Education ›
Volume 43, 2012 - Issue 2

ECONOMIC INSTRUCTION

Teaching Integrity in Empirical Research: A Protocol for Documenting Data Management and Analysis

Richard Ball & Norm Medeiros

Pages 182-189 | Published online: 11 Apr 2012
Download citation https://doi.org/10.1080/00220485.2012.659647

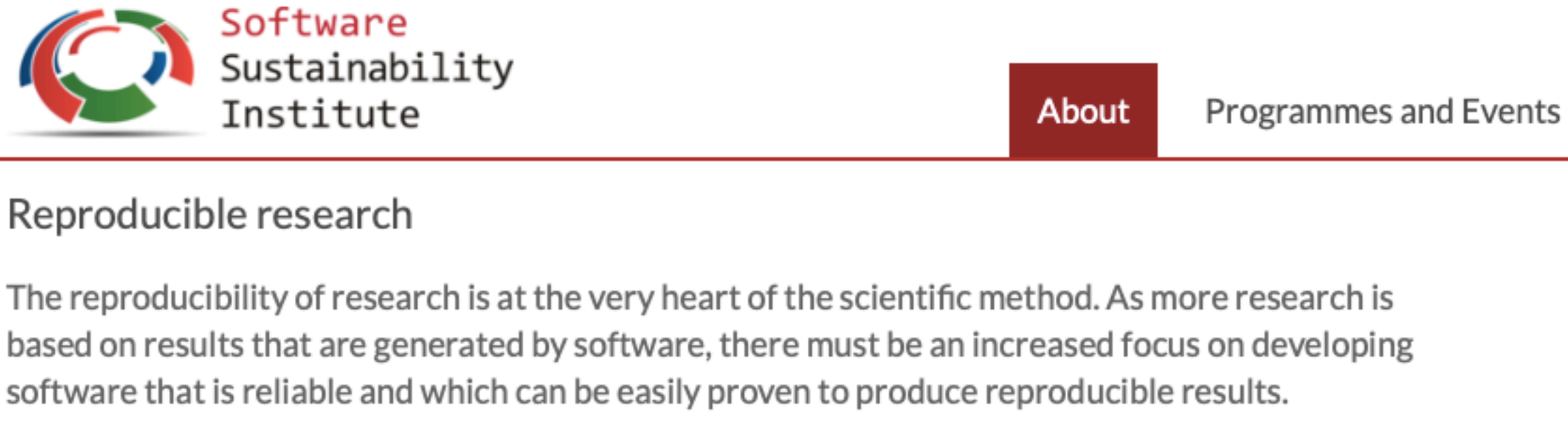
Abstract

This article describes a protocol the authors developed for teaching undergraduates to document their statistical analyses for empirical research projects so that their results are completely reproducible and verifiable. The protocol is guided by the principle that the documentation prepared to accompany an empirical research project should be sufficient to allow an independent researcher to replicate easily and exactly every step of the data management and analysis that generated the results reported in a study. The authors hope that requiring students to follow this protocol will not only teach them how to document their research appropriately, but also instill in them the belief that such documentation is an important professional responsibility.

Keywords: documentation, empirical research, replication

https://www.tandfonline.com/doi/abs/10.1080/00220485.2012.659647

The Software Sustainability Institute



https://www.software.ac.uk/about/manifesto

Lots of Open Science-related talks and activities in the pipeline incl. Lancaster June 4 for PhD students, RUM workshop on using Binder to reproduce your computational environment (June 12), CarpentryConnect workshop Manchester, June 25/26/27.

Dr Kirstie Whittaker
(University of Cambridge)
The Turing Way

Dr Lisa de Bruine
(University of Glasgow)
Large-scale collaboration and
the Psychological Science Accelerator

Prof Chris Chambers
(Cardiff University)
Q&A on Registered Reports
journal submissions

Dr Andrew Stewart
(University of Manchester)
Reproducible Data Visualization

Other topics include:

Dealing with Big Data
Study Pre-registration
The Many Babies Project
Publishing and Open Research

Software Sustainability Institute

Future-proofing Your Research

Moving Towards Open and Reproducible Research Practices

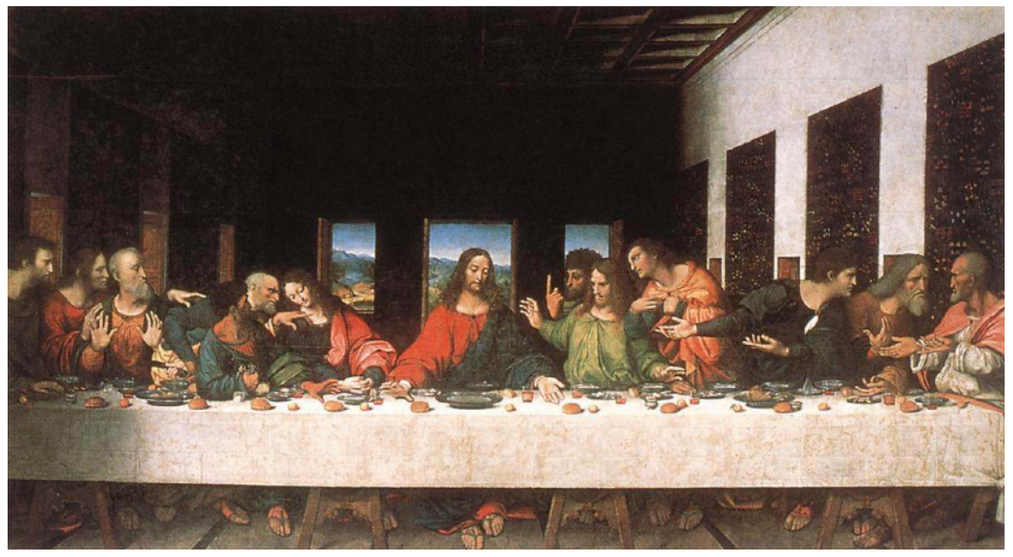
4th June 2019 10:00-16:00
(Coffee and registration 09:30)

Lancaster University Library

Register for free
bit.ly/OpenRes

METHODS NORTH WEST | MANCHESTER 1824 The University of Manchester | Lancaster University | PROSPR

Now is a HUGELY exciting time to be working as a psychologist - we are all part of a renaissance of the methods we use to conduct, analyse, and report psychological research…



Thank You!

andrew.stewart@manchester.ac.uk

@ajstewart_lang

https://listserv.manchester.ac.uk/cgi-bin/wa?REPORT=OPEN_RESEARCH

Software Sustainability Institute

Slides here:
https://ajstewartlang.github.io/Keele_Staffs_talk.pdf