



# Lecture 05: The Open Science movement in Psychology

*Doing better*

Dr. Gordon Wright 

*g.wright@gold.ac.uk*

October 30, 2023



# Key topics today

The week ahead (week 5)

- Personal Tutor Meeting about essay writing - bring your questions
- Design & Analysis Quiz due this week (week 5)
- Open Science
- Labs - Critical Proposal and Power Calculations



# Personal Tutor Meeting Week 5

This week (week 5) your PT session is all about essay writing



Some of you have expressed doubts about this. Please see this as an opportunity to get answers to any questions.

Make sure to use your feedback!



# Any Questions?



So...



But what does that mean?



# Open Science Collaboration (2015)

## RESEARCH ARTICLE SUMMARY

### PSYCHOLOGY

# Estimating the reproducibility of psychological science

Open Science Collaboration\*

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), 943–943.  
<http://www.jstor.org/stable/24749235>



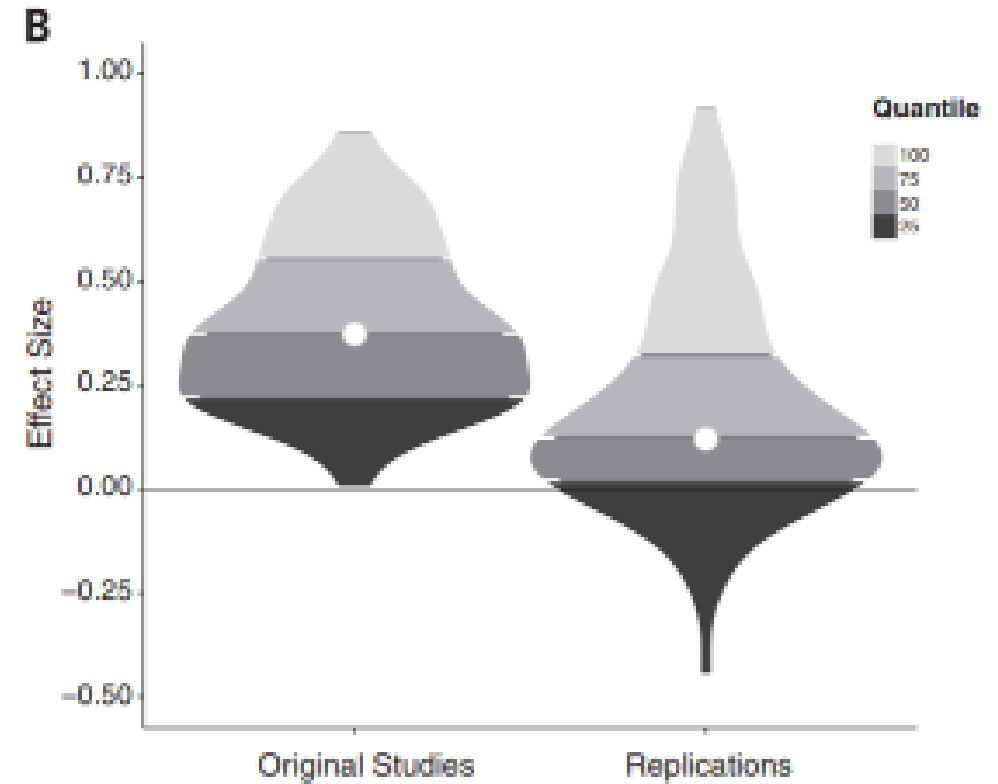
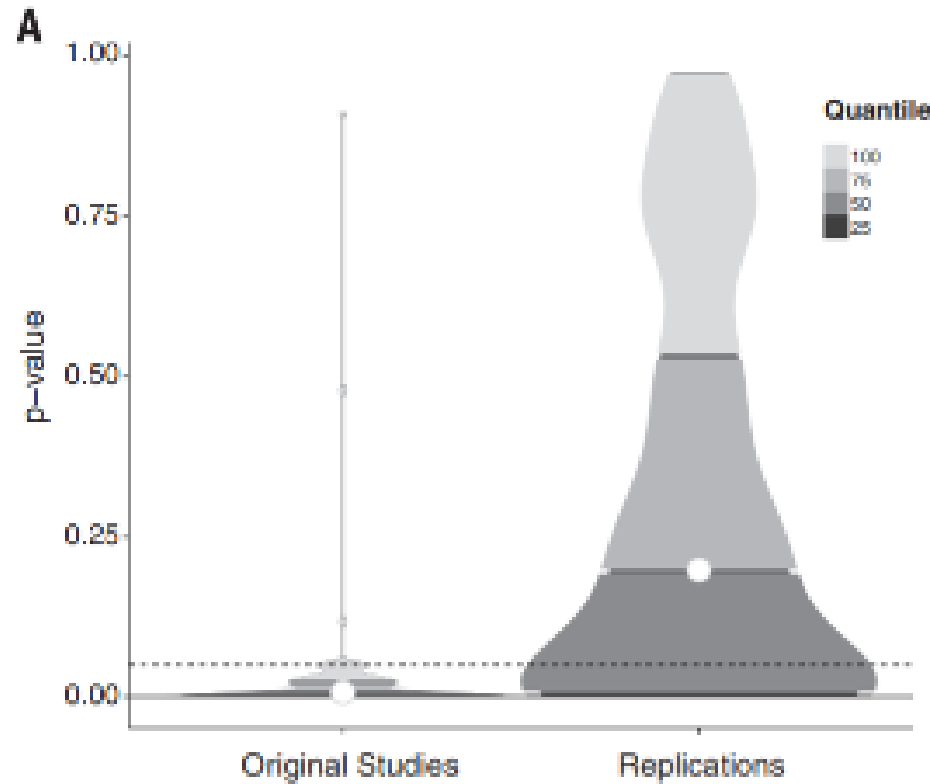
# The replication crisis

- The Open Science Collaboration (2015) (c.f. Brian Nosek) conducted 100 replications of psychology studies published in three psychology journals
- While 97 of previous studies reported significant results, only 36 were significant in the replication attempt. And effects were smaller than originally reported...





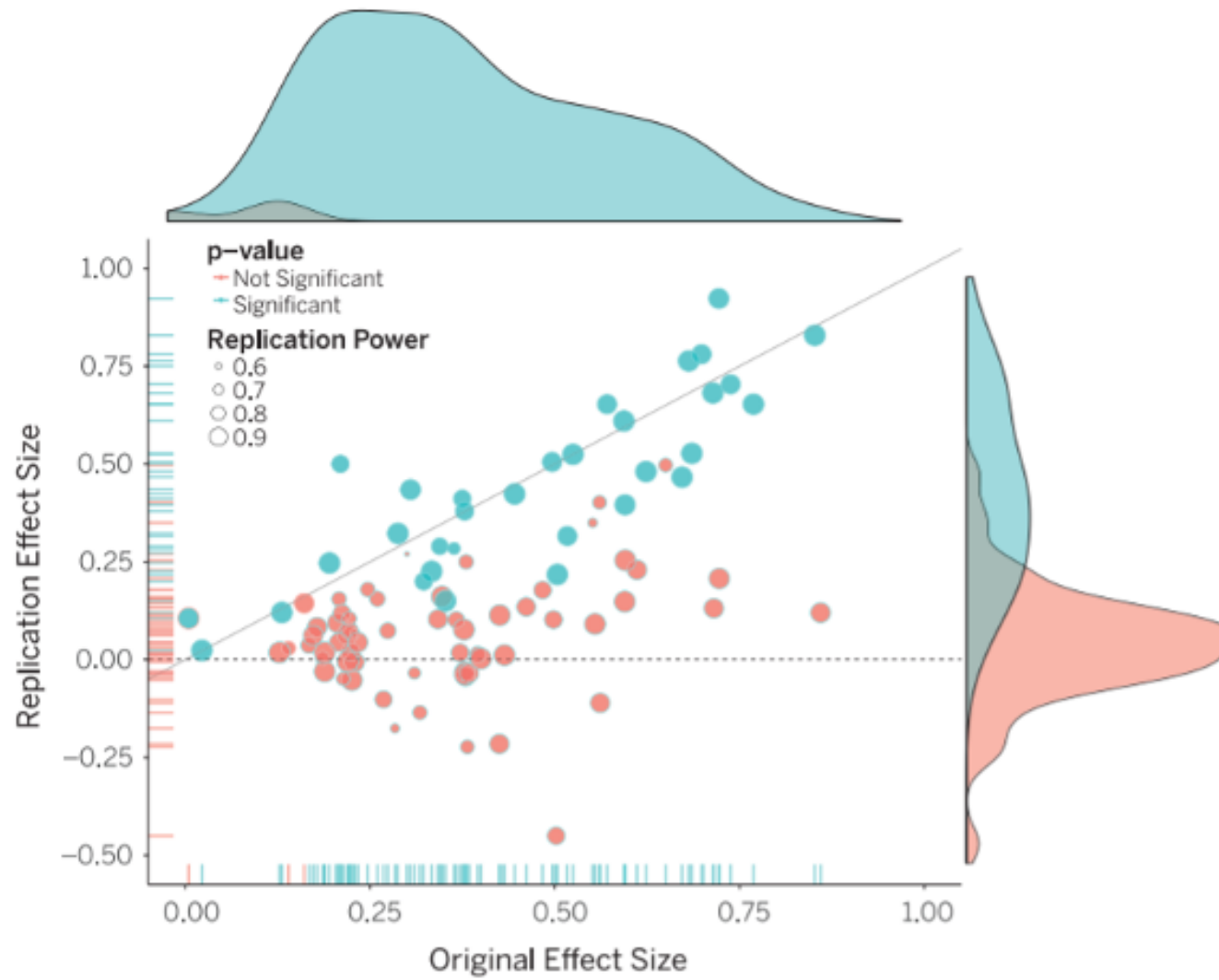
# Violin plots



Violin Plots of Replication Results



# Raincloud plots



Raincloud Plots of Replication Results



# Why aren't we replicating?

- Some point the finger at scientific fraud (i.e. bad scientists making up their data)
- However, others point to more systematic problems
- Low statistical power
- Questionable research practices (QRPs)
- Publication bias



# Statistical power

- Since 1960s, sample sizes in standard psychology studies have remained too small – giving them low power
- Low power is normally a problem because it means that you don't find significant effects
- An underappreciated downside of low power is that if you do find effect, it is probably spuriously exaggerated
- This will mean that when you try to replicate it, it will be smaller (not significant)



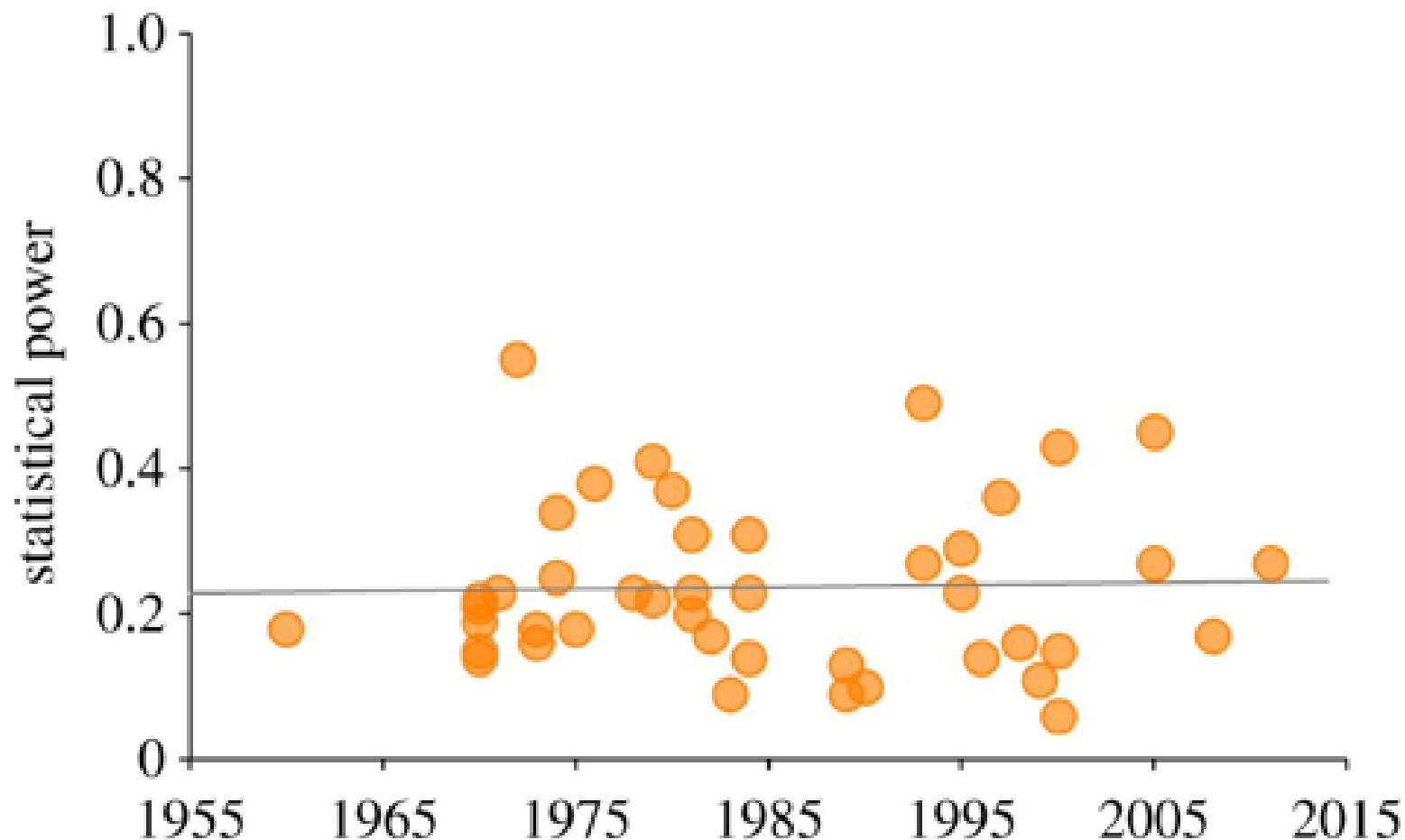
# We are training you in best practice

If you have had trouble finding an effect size in your Personality Essay or Critical Proposal...

This is either because the new best practice hasn't been adopted, or the research team dropped the ball.



# Power plot



Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9), 160384.

<https://doi.org/10.1098/rsos.160384>



# Power and Power Calculations in Psychology



# What is power?

- Power is the probability of rejecting the null hypothesis when it is false.
- Power depends on the significance level, sample size, and effect size of a test.
- Power is important for planning and evaluating studies.





# How to calculate power?

- Use online tools or statistical software like G\*Power.
- Specify the type of test, the alpha level, the effect size, and the desired power or sample size.
- For complex research designs, you may need to calculate a number of potential effect sizes



# Why is power low in psychology?

- Small sample sizes are common in psychological research.
- Effect sizes are often unknown or overestimated.
- Researchers may not use power analysis or understand its meaning.



# How to improve power in psychology?

- Increase sample size or use more sensitive measures.
- Use meta-analysis or replication to estimate effect sizes.
- Educate researchers and reviewers about power and its implications.



# Questionable Research Practices (QRPs)

## **Selective reporting of participants**

E.g., excluding data from some participants

## **Selective reporting of manipulations or variables**

E.g., measuring many different variables in a study, but only writing up the variables that ‘worked’ (were significant)

## **Optional stopping rules**

E.g., continuing to add participants to a sample until it is just significant ( $p < .05$ )



## QRPs Continued

### Flexible data analysis

E.g., Adding covariates (without good reason) to ‘improve’ statistical results

### HARKing (Hypothesising After Results are Known)

Running a study, and then generating a hypothesis that fits the results (even if they were not what you originally predicted)

**What these practices all have in common is they involve capitalising on chance to create a significant result (which may not be reliable)**

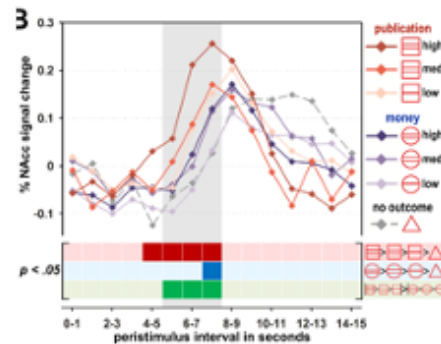
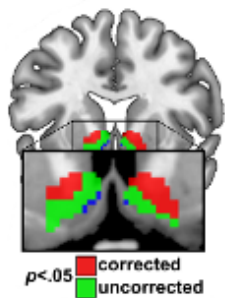
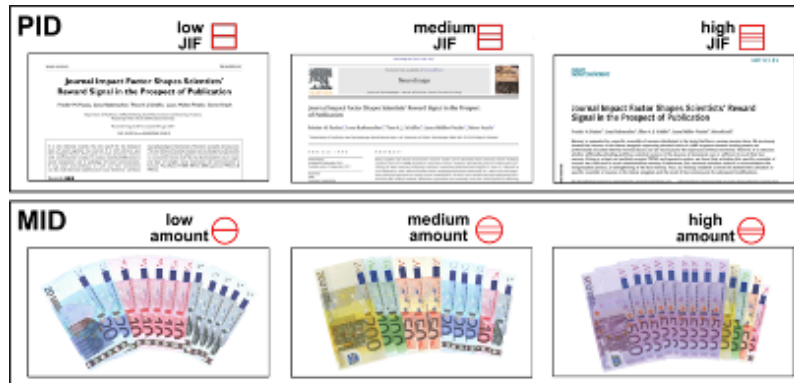


# Novelty and glamour

- Scientists want to communicate their science, but they also want successful careers
- An important metric for success in science is publishing in ‘top journals’ (e.g., Nature, Science)
- Getting published in these journals gets your science out to a wide audience (because lots of people read them) but also carries prestige – you get jobs, grants, funding and prizes from publishing regularly in these journals
- But top journals want to publish novel or surprising results.
- Why do you think that could be a problem?



# Lust for Impact Factors!



Paulus, F. M., Rademacher, L., Schäfer, T. A., Müller-Pinzler, L., & Krach, S. (2015). Journal Impact Factor Shapes Scientists' Reward Signal in the Prospect of Publication. *PloS one*, 10(11), e0142537. <https://doi.org/10.1371/journal.pone.0142537>



# Biases in journals: File drawer problem

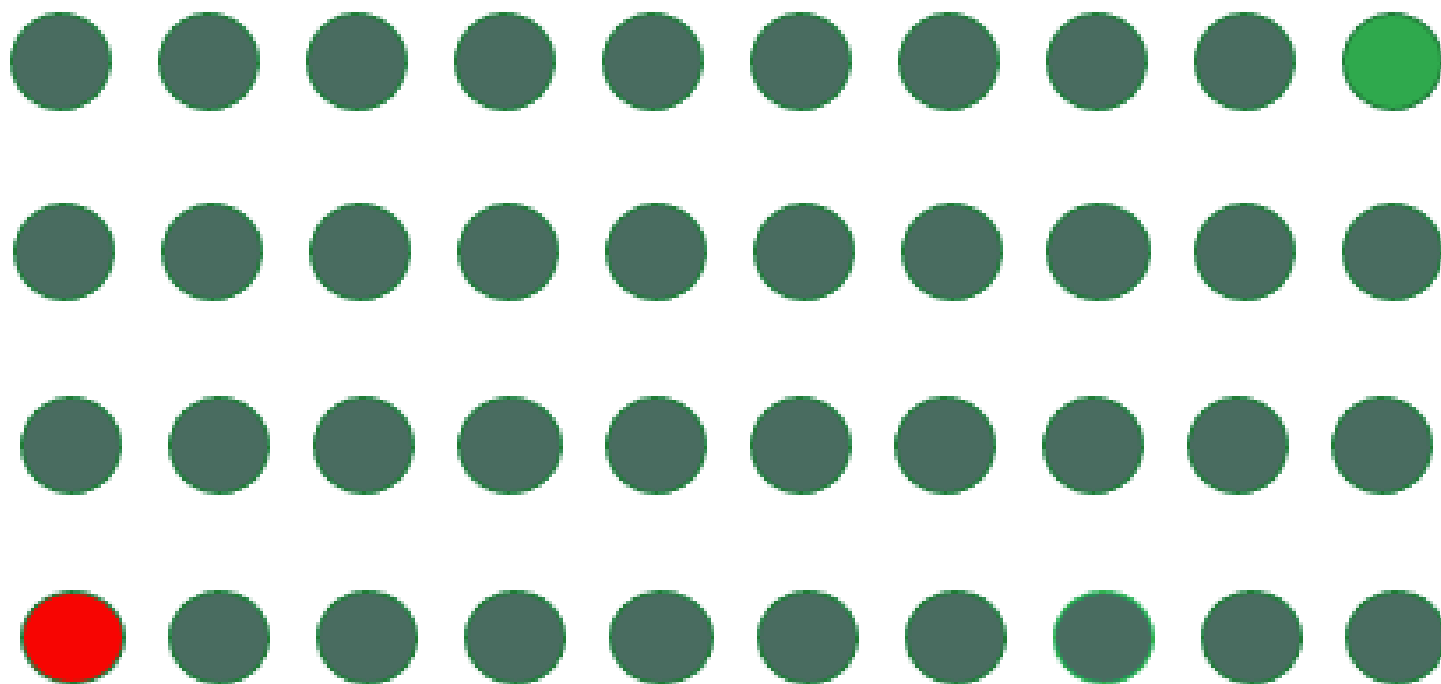
- Even beyond ‘prestige’ journals, journals are biased to publish positive (i.e. significant) findings
- Because it is much easier to publish positive results, rather than nonsignificant results or failed replications, science has a ‘file drawer problem’
- Scientists don’t try to publish their null results, and/or journals make it hard to publish them
- This means the published literature is biased to contain significant results (that come from a distribution where there is no true effect)





## Let's work the probabilities

With an alpha level of  $p=.05$ , if we have 40 scientists testing any hypothesis we would expect one to find a significant result in one direction, and another to find a significant result in another direction just by random chance





# The credibility revolution?

Recent years have seen several changes to how psychological science is conducted to overcome concerns about reliability – dubbed the ‘credibility revolution’

## Implications of the Credibility Revolution for Productivity, Creativity, and Progress

**Simine Vazire**

Department of Psychology, University of California, Davis

### Abstract

The credibility revolution (sometimes referred to as the “replicability crisis”) in psychology has brought about many changes in the standards by which psychological science is evaluated. These changes include (a) greater emphasis on transparency and openness, (b) a move toward preregistration of research, (c) more direct-replication studies, and (d) higher standards for the quality and quantity of evidence needed to make strong scientific claims. What are the implications of these changes for productivity, creativity, and progress in psychological science? These questions can and should be studied empirically, and I present my predictions here. The productivity of individual researchers is likely to decline, although some changes (e.g., greater collaboration, data sharing) may mitigate this effect. The effects of these changes on creativity are likely to be mixed: Researchers will be less likely to pursue risky questions; more likely to use a broad range of methods, designs, and populations; and less free to define their own best practices and standards of evidence. Finally, the rate of scientific progress—the most important shared goal of scientists—is likely to increase as a result of these changes, although one’s subjective experience of making progress will likely become rarer.

*perspectives on psychological science*  
2018, Vol. 13(4) 411–417  
© The Author(s) 2018  
Reprints and permissions:  
[sagepub.com/journalsPermissions.nav](http://sagepub.com/journalsPermissions.nav)  
DOI: 10.1177/1745691617751884  
[www.psychologicalscience.org/PPS](http://www.psychologicalscience.org/PPS)  
 SAGE



# Recommendations and changes

Low statistical power? Report power analyses and justify sample sizes

Method and Results: These sections of Research Articles do not count toward the total word limit. The aim of unrestricted length for Method and Results sections is to allow authors to provide clear, complete, self-contained descriptions of their studies. But as much as Psychological Science prizes narrative clarity and completeness, so too does it value concision. In almost all cases, an adequate account of method and results can be achieved in 2,500 or fewer words for Research Articles. **Methodological minutiae and fine-grained details on the Results—the sorts of information that only “insiders” would relish and require for purposes of replication—should be placed in Supplemental Online Materials-Reviewed, not in the main text.** Authors should include in their Method sections (a) justification for the sample(s) selected for the study (if the sample is of convenience, this should be explicitly noted); (b) the total number of excluded observations and the reasons for making the exclusions (if any); and (c) an explanation as to why the sample size is considered reasonable, supported by a formal power analysis, if appropriate. Authors also should include confirmation in their Method section that the research meets relevant ethical guidelines, including adherence to the legal requirements of the study country

(Taken from guidance to authors at journal Psychological Science)



# Familiar?



## A Multisite Preregistered Paradigmatic Test of the Ego-Depletion Effect

Kathleen D. Vohs<sup>ID</sup>, Brandon J. Schmelchel, Sophie Lohmann<sup>ID</sup>, Quentin F. Gronau<sup>ID</sup>, Anna J. Finley, Sarah E. Ainsworth, Jessica L. Alquist, Michael D. Baker, Ambra Brizi, Angelica Bunyi, Grant J. Butschek, Collier Campbell, Jonathan Capaldi, Chuting Cau, Heather Chambers, Nikos L. D. Chatzisarantis, Weston J. Christensen, Samuel L. Clay, Jessica Curtis, Valeria De Cristofaro, Kareena del Rosario, Katharina Diel, Yasemin Doğruol, Megan Dol, Tina L. Donaldson, Andreas B. Eder, Mia Ersoff, Julie R. Eyink, Angelica Falkenstein, Bob M. Fennis, Matthew B. Findley, Eli J. Finkel, Victoria Forgea, Malte Friese, Paul Fuglestad, Natasha E. Garcia-Willingham, Lea F. Geraedts, Will M. Gervais, Mauro Giacomantonio, Bryan Gibson, Karolin Gieseler, Justina Ginelkiene, Elana M. Gloger, Carina M. Gobes, Maria Grande, Martin S. Hagger, Bethany Hartsell, Anthony D. Hermann, Jasper J. Hidding, Edward R. Hirt, Josh Hodge, Wilhelm Hofmann, Jennifer L. Howell, Robert D. Hutton, Michael Inzlicht, Lily James, Emily Johnson, Hannah L. Johnson, Sarah M. Joyce, Yannick Joye, Jan Helge Kaben, Lara K. Kammrath, Caitlin N. Kelly, Brian L. Kissell, Sander L. Koole, Anand Krishna, Christine Lam, Kelemen T. Lee, Nick Lee, Dana C. Leighton, David D. Loschelder, Heather M. Maranges, E. J. Masicampo, Kennedy Mazara, Jr., Samantha McCarthy, Ian McGregor, Nicole L. Mead, Wendy B. Mendes, Carine Meslot, Nicholas M. Michalak, Marina Milyavskaya, Akira Miyake, Mehrad Moeini-Jazani, Mark Muraven, Erin Nakahara, Krishna Patel, John V. Petrocelli, Katja M. Pollak, Mindi M. Price, Haley J. Ramsey, Maximilian Rath, Jacob A. Robertson, Rachael Rockwell, Isabella F. Russ, Marco Salvati, Blair Saunders, Anne Scherer, Astrid Schütz, Kristin N. Schmitt, Suzanne C. Segerstrom, Benjamin Serenka, Konstantyn Sharpinskyi, Meaghan Shaw, Janelle Sherman, Yu Song, Nicholas Sosa, Kaitlyn Spillane, Julia Stapels, Alec J. Stinnett, Hannah R. Strawser, Kate Sweeny, Dominic Theodore, Karine Tonnu, Yasmijn van Oldenbeuving, Michelle R. vanDellen, Raiza C. Vergara, Jasmine S. Walker, Christian E. Waugh, Feline Weise, Kaitlyn M. Werner, Craig Wheeler, Rachel A. White, Aaron L. Wichman, Bradford J. Wiggins, Julian A. Wills, Janie H. Wilson, Eric-Jan Wagenmakers<sup>ID</sup>, Dolores Albarracín

First Published September 14, 2021; pp. 1566–1581

[Abstract](#)

[> Preview](#)





# The goal



## Psychological Drivers of Individual Differences in Risk Perception: A Systematic Case Study Focusing on 5G

Renato Frey

First Published September 22, 2021; pp. 1592–1604

[Abstract](#)

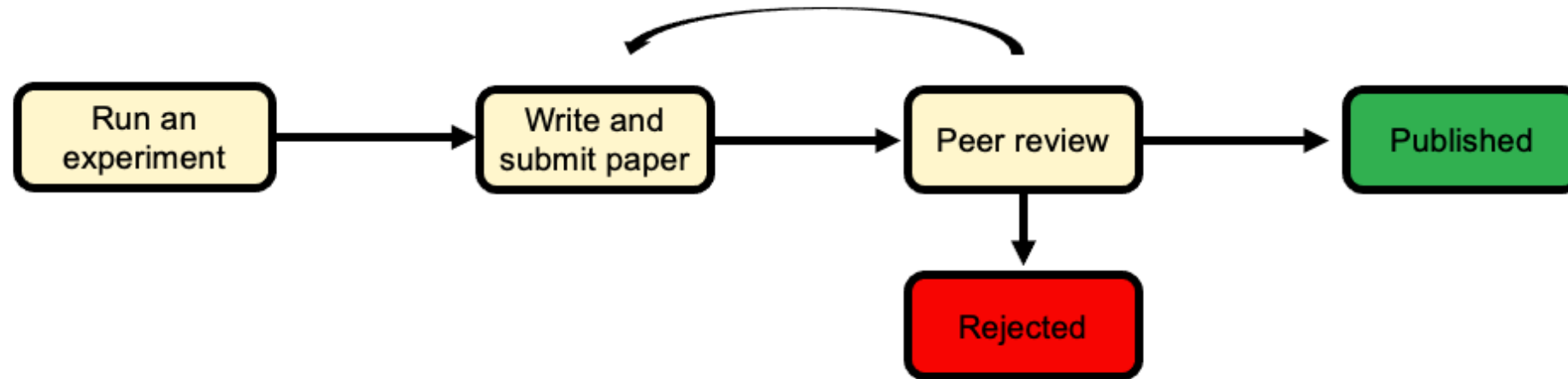
[> Preview](#)





# The 'normal' process

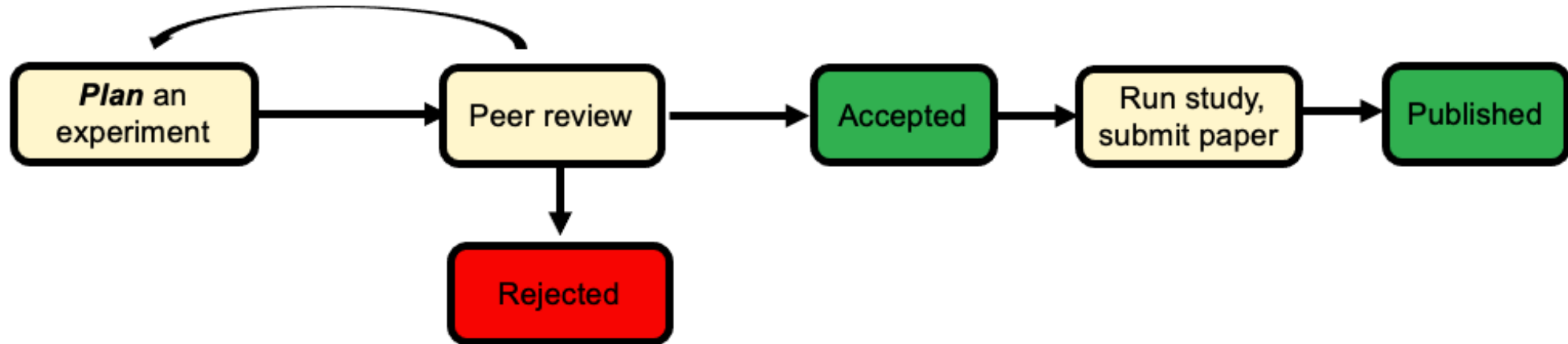
*Normal peer review process*





# A better solution?

*Registered report*





# Do scientists already ‘know’ which results to trust?


- The unnerving thing about the ‘replication crisis’ seems to be that psychological theories are built on foundations of sand. But is this true?
- Camerer and colleagues attempted to replicate 21 social science studies (including psychology) and found around 13 replicated.
- However, the study also ran a prediction market where scientists (PhD or PhD student) had to bet on which studies would replicate and which wouldn’t
- We should want our journal to publish things that are robust – but if scientists have a good sense of what is reliable, is this really a ‘crisis’?





Camerer et al. (2018)

## Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

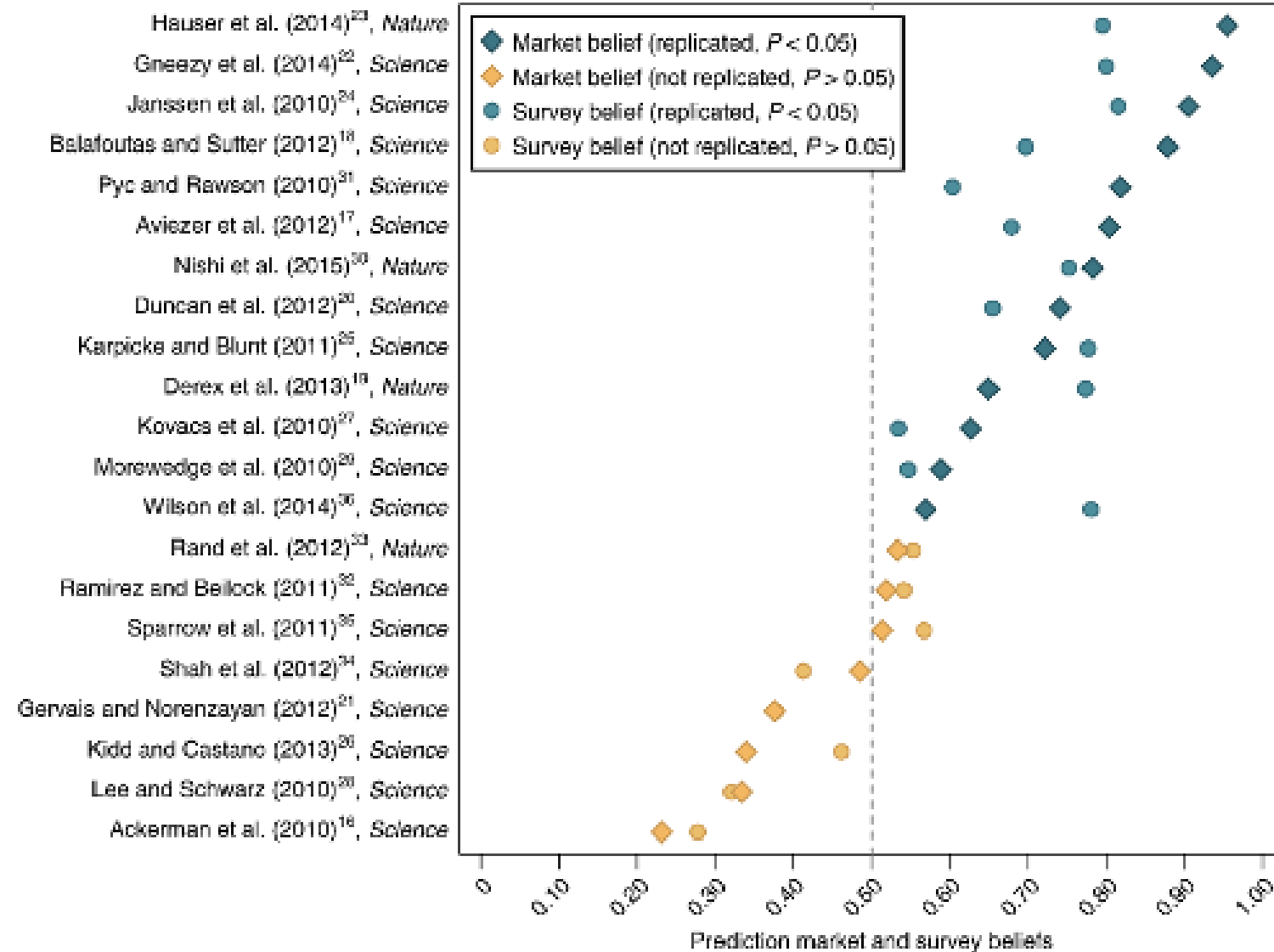
[Colin F. Camerer](#), [Anna Dreber](#), [Felix Holzmeister](#), [Teck-Hua Ho](#), [Jürgen Huber](#), [Magnus Johannesson](#), [Michael Kirchler](#), [Gideon Nave](#), [Brian A. Nosek](#) , [Thomas Pfeiffer](#), [Adam Altmejd](#), [Nick Buttrick](#), [Taizan Chan](#), [Yiling Chen](#), [Eskil Forsell](#), [Anup Gampa](#), [Emma Heikensten](#), [Lily Hummer](#), [Taisuke Imai](#), [Siri Isaksson](#), [Dylan Manfredi](#), [Julia Rose](#), [Eric-Jan Wagenmakers](#) & [Hang Wu](#)

[Nature Human Behaviour](#) **2**, 637–644 (2018) | [Cite this article](#)

**60k** Accesses | **544** Citations | **2338** Altmetric | [Metrics](#)



# Findings





# Dubious efforts to replicate

Researchers who do replication studies also have flexibility in their design and analysis choices.

There may be a bias to not replicate certain findings (e.g., because you are sceptical of the result in the first place)

## Replicator degrees of freedom allow publication of misleading failures to replicate

Christopher J. Bryan<sup>a,1</sup>, David S. Yeager<sup>b</sup>, and Joseph M. O'Brien<sup>b</sup>

<sup>a</sup>Booth School of Business, University of Chicago, Chicago, IL 60637; and <sup>b</sup>Department of Psychology, University of Texas at Austin, Austin, TX 78712

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved October 22, 2019 (received for review June 28, 2019)

In recent years, the field of psychology has begun to conduct replication tests on a large scale. Here, we show that “replicator degrees of freedom” make it far too easy to obtain and publish false-negative replication results, even while appearing to adhere to strict methodological standards. Specifically, using data from an ongoing debate, we show that commonly exercised flexibility at the experimental design and data analysis stages of replication testing can make it appear that a finding was not replicated when, in fact, it was. The debate that we focus on is representative, on key dimensions, of a large number of other replication tests in psychology that have been published in recent years, suggesting that the lessons of this analysis may be far reaching. The problems with current practice in replication science that we uncover here are particularly worrisome because they are not adequately addressed by the field’s standard remedies, including preregistration. Implications for how the field could develop more effective methodological standards for replication are discussed.

they could have an ironic and counterproductive effect: trading one sort of misleading research finding (false-positive original findings) for another (false-negative replication results). This is a bad trade because the latter sort of misleading finding undoes the field’s hard-won progress toward improved scientific understanding.

Others have already made versions of the 2 general methodological points that we make here: that empirical conclusions often hinge on analytic choices that competent investigators can disagree about and that replication tests that deviate from the design of the original study in material ways can create the misleading impression that the original finding was a false positive (19–25). Here, we provide an analysis of one prominent ongoing replication debate that demonstrates, concretely and directly, the implications of these 2 methodological principles for the field’s interpretation of the many ostensible failures to replicate that are already in the literature and for how replication tests should be conducted going forward.

The failure of many replication tests to adequately represent



## No reason to worry

Some have suggested that low replication rates are not necessarily a sign of bad research

Alexander Bird (philosopher of science) suggests worries about replication reflect base rate fallacy

Most hypotheses are wrong so we wouldn't expect them to replicate in future studies

What do you think?



# Alexander Bird (2018)

## Understanding the Replication Crisis as a Base Rate Fallacy

Alexander Bird 

---

### ABSTRACT

The replication (replicability, reproducibility) crisis in social psychology and clinical medicine arises from the fact that many apparently well-confirmed experimental results are subsequently overturned by studies that aim to replicate the original study. The culprit is widely held to be poor science: questionable research practices, failure to publish negative results, bad incentives, and even fraud. In this article I argue that the high rate of failed replications is consistent with high-quality science. **We would expect this outcome if the field of science in question produces a high proportion of false hypotheses prior to testing.** If most of the hypotheses under test are false, then there will be many false hypotheses that are apparently supported by the outcomes of well conducted experiments and null hypothesis significance tests with a type-I error rate ( $\alpha$ ) of 5%. Failure to recognize this is to commit the fallacy of ignoring the base rate. I argue that this is a plausible diagnosis of the replication crisis and examine what lessons we thereby learn for the future conduct of science.



# Are we worry about the wrong thing?

- Other psychologists have argued that focus on replicability, statistical robustness etc. is misguided
- The real problem psychology has is the absence of strong theories
- This “theory crisis” cannot be solved with more and more attention to statistics
- Theory is the thing we should be caring about? Not specific effects in specific studies
- No statistics can help us to test a theory that is poorly thought out



# Summary

You should now know:

- Why scientists are concerned about the reliability of psychological studies
- Steps the scientific community are taking to overcome these worries
- Not everyone is convinced that the ‘crisis’ is as serious as it seems, or whether these changes will help solve psychology’s problems



# Questions?

## Lab activities

Power Calculations in preparation for your Ethics Applications

Pay close attention to the lab slides - Step by Step guidance for EVERY ANOVA flavour

Access to detailed Ethics VLE page and resources (Please Review)

Lab Priority is the Critical Proposal Target Paper approval/confirmation





# References

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>

