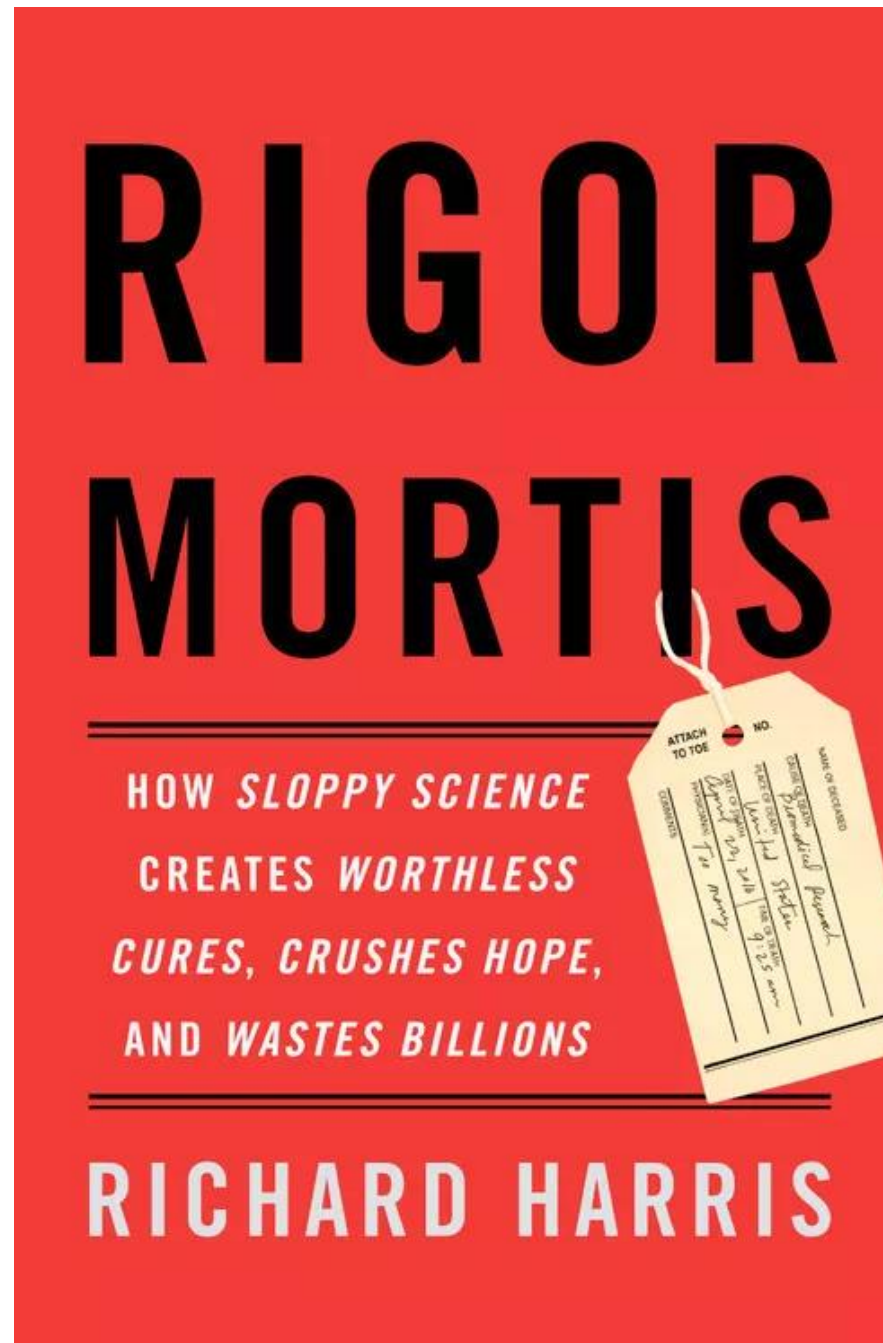


The Replication Crisis in Science: An Overview and Current Results

Dr. Axel Kohler
Goethe Research Academy for Early Career Researchers (GRADE)

GU Open Science Forum 2022
June 30, 2022

The Costs of Sloppy Science



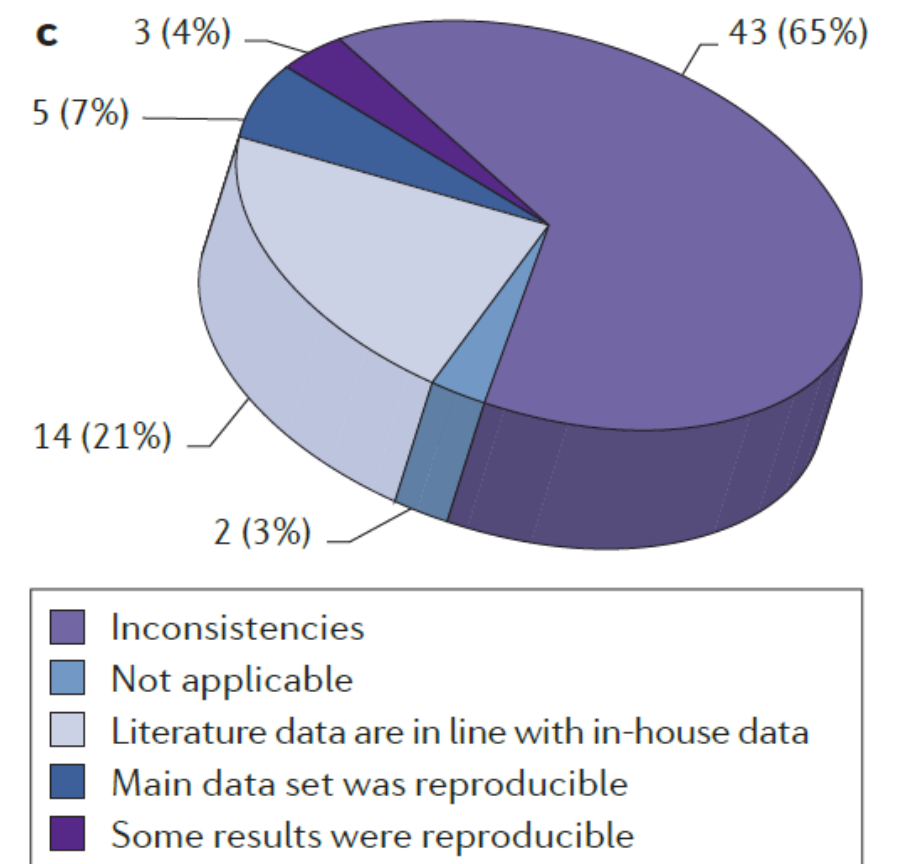
“American taxpayers spend \$30 billion annually funding biomedical research. By some estimates, half of the results from these studies can't be replicated elsewhere—the science is simply wrong. **Often, research institutes and academia emphasize publishing results over getting the right answers, incentivizing poor experimental design, improper methods, and sloppy statistics.**”

From the book announcement

Negative Impact on Science and Society

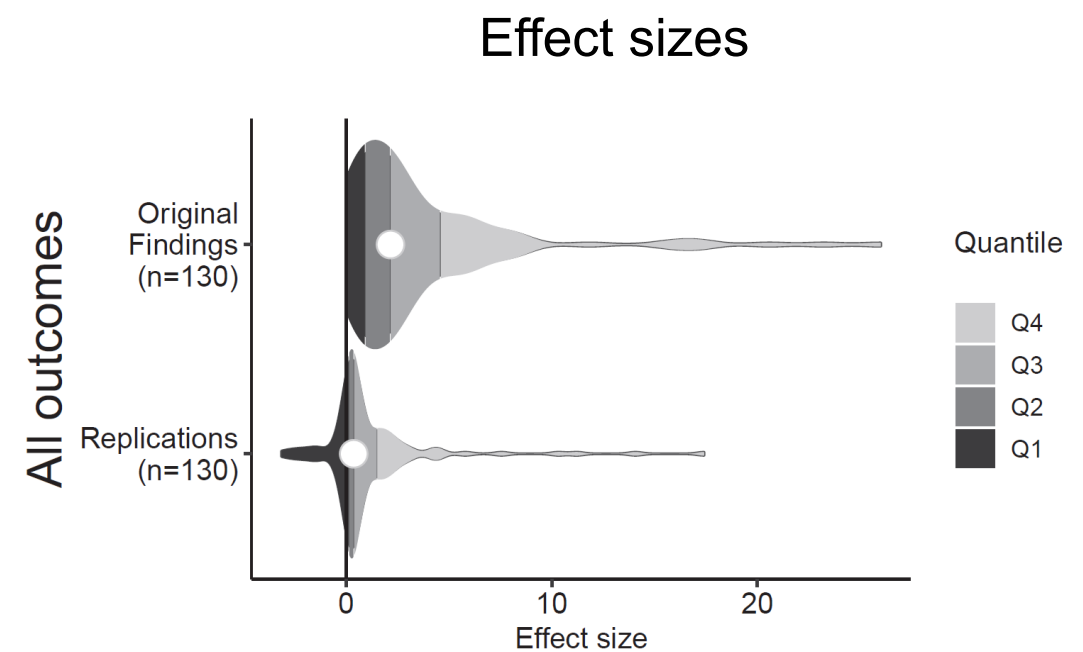
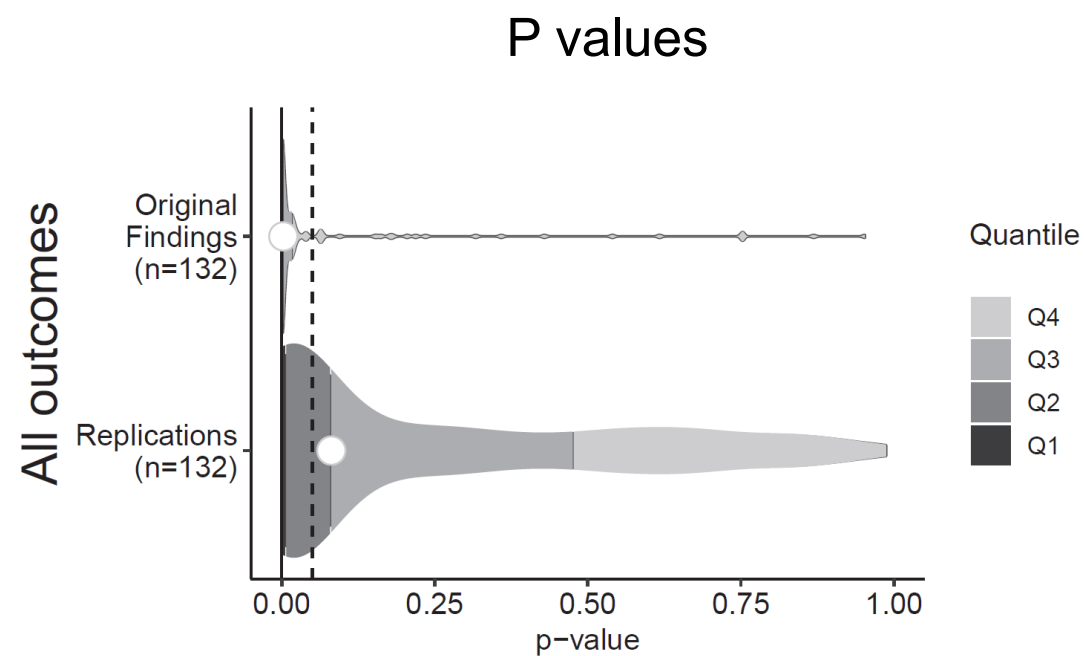
- Time, effort, and resources are wasted building on previous results that are not reproducible
- Trust in science is eroding and possibly weakens impact on politics
- Drug companies have increasing reservations concerning scientific results

literature vs. company data



Replication Project in Cancer Biology

- Replication of **50 experiments** from high-impact journals in preclinical cancer research
- **Replication success of 40%** for significant effects (80% for null effects)
- **Effect size about 22%** of the original effect size
- **Side note:** The originally planned 150 experiments could not be performed due to problems with the documentation of the experiments



Replication Project in Psychology

- Replication of **100 experiments** from high-impact by a collective of **270 authors**
- **Replication success of 37%** (confirmed significance; most replications had larger samples)
- **Effect size about 50%** of the original effect size

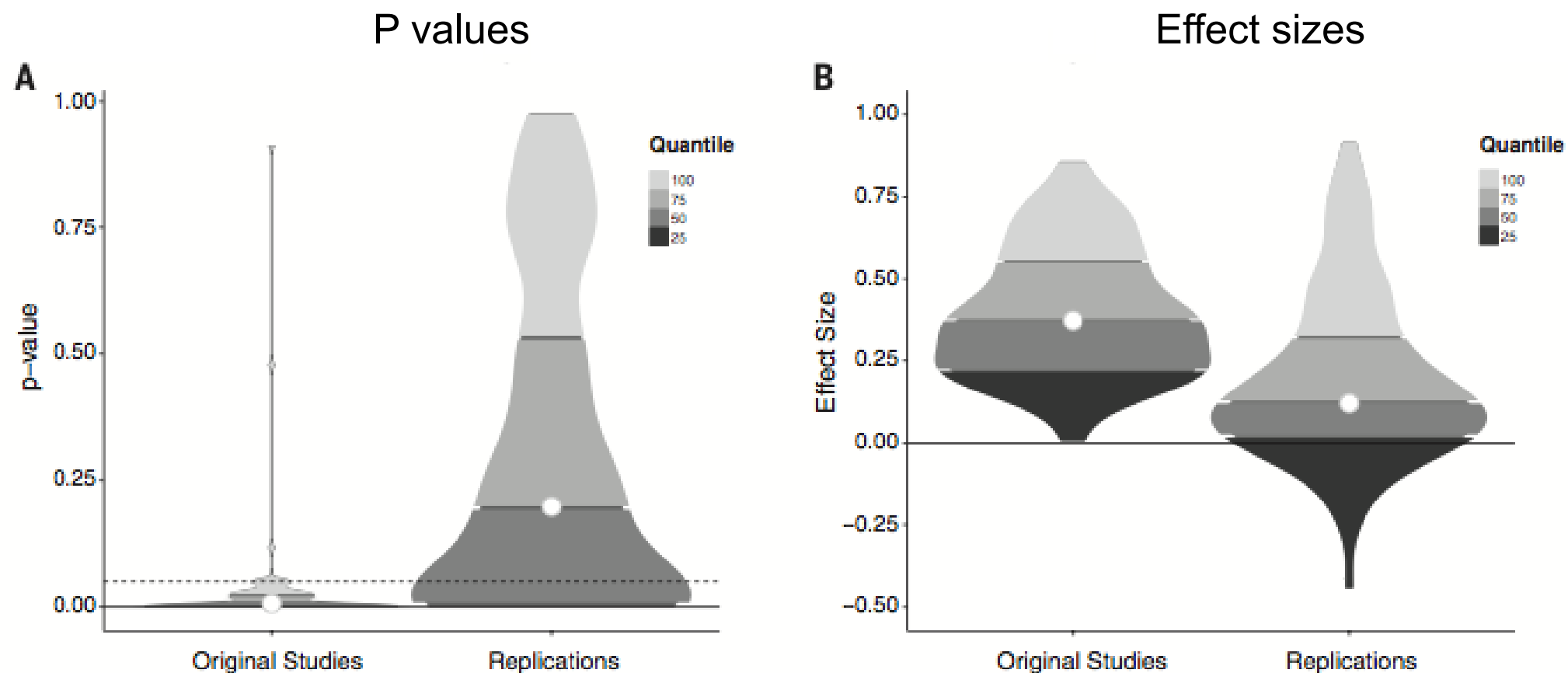
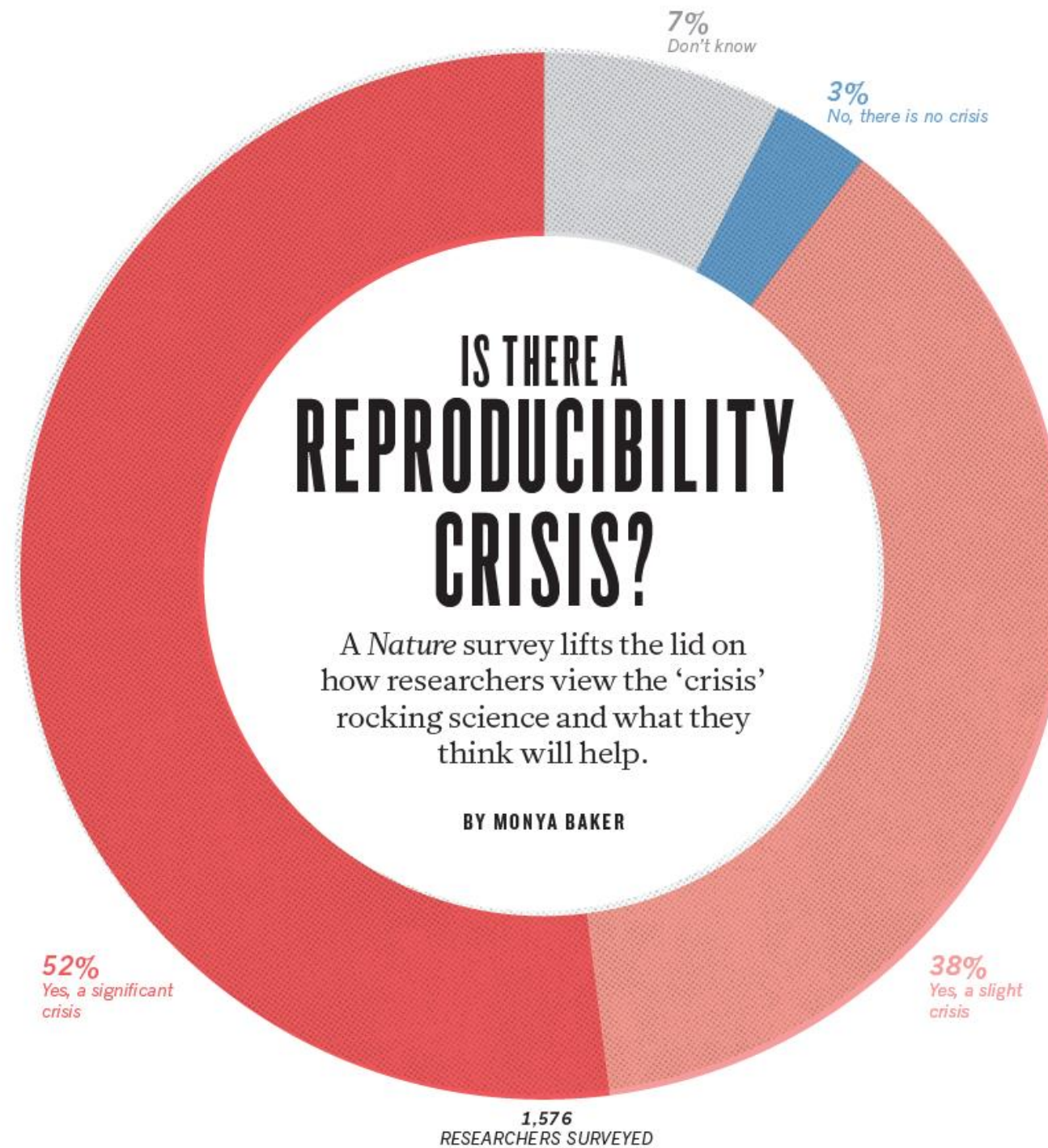


Fig. 1. Density plots of original and replication P values and effect sizes. (A) P values. (B) Effect sizes (correlation coefficients). Lowest quantiles for P values are not visible because they are clustered near zero.

Reproducibility Crisis in Other Fields



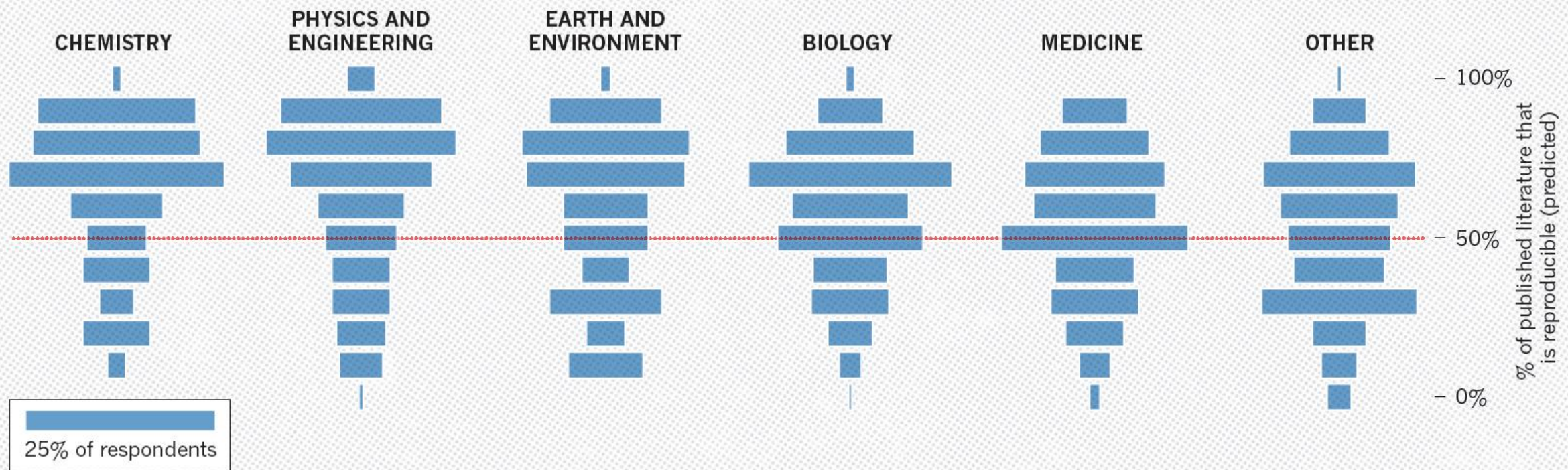
Reproducibility Crisis in Other Fields

A 'CRISIS' IN NUMBERS

Nature surveyed 1,576 scientists online to get their thoughts on reproducibility in their field and in science in general. See go.nature.com/2vjr4y for more charts and access to the full data.

HOW MUCH PUBLISHED WORK IN YOUR FIELD IS REPRODUCIBLE?

Physicists and chemists were most confident in the literature.



Improbable Results in Psychology

Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

© 2011 American Psychological Association
0022-3514/11/\$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of *psi* are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are

**!!EXTRAORDINARY CLAIMS REQUIRE
EXTRAORDINARY EVIDENCE!!**

Rebuttal by Wagenmakers et al. (2011, JPSP):

„We conclude that Bem’s *p* values do not indicate evidence in favor of precognition; instead, they indicate that **experimental psychologists need to change the way they conduct their experiments and analyze their data.**“

The Problem – Researcher Degrees of Freedom

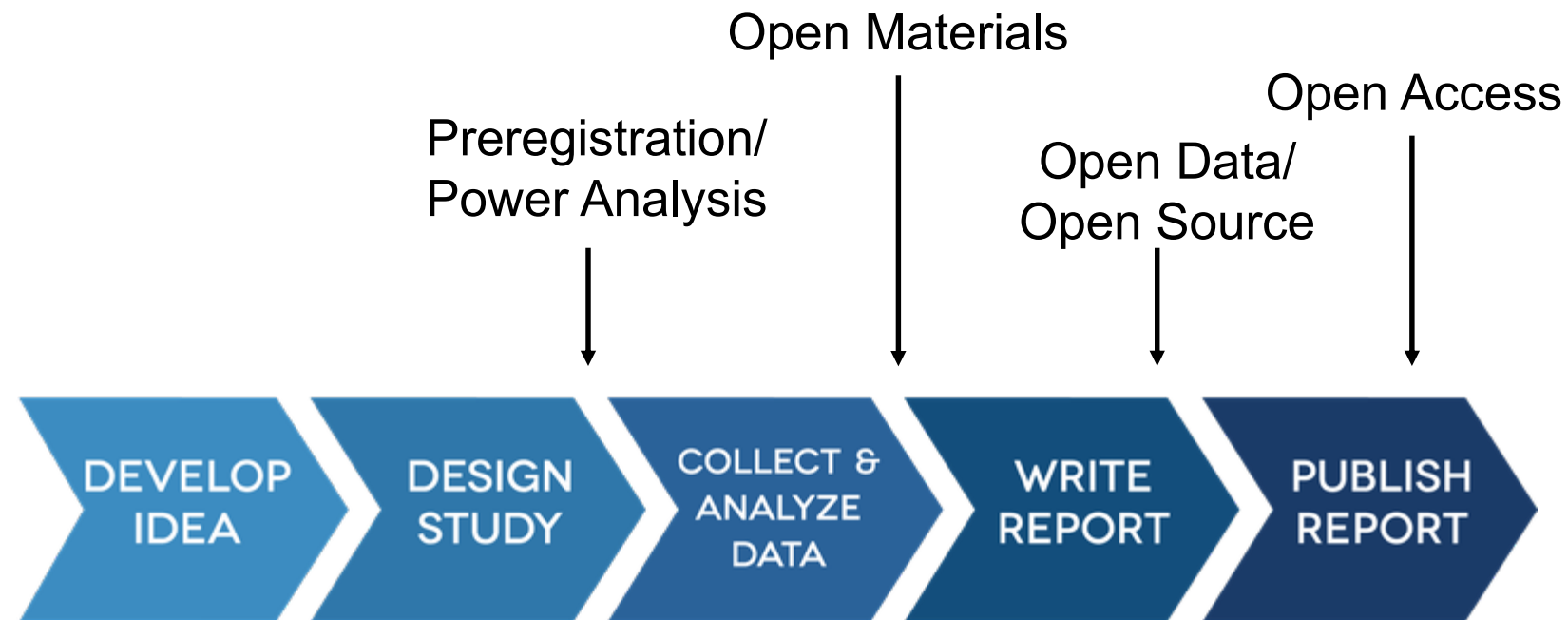
- **Using flexible sampling rules:** Should more data be collected?
- **Selecting observations:** Should some observations be excluded?
- **Using multiple independent variables/conditions:** Which conditions should be combined and which ones compared?
- **Adding and leaving out covariates:** Which control variables should be considered?
- **Using multiple dependent variables/effect measures:** Should specific measures be combined or transformed or both?

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. <https://doi.org/10.1177/0956797611417632>

(Other) Questionable Research Practices

- **p-Hacking:** Doing everything to get that damn p-value
- **HARKing:** Hypothesizing After Results are Known – quite common distortion of assumptions held a priori
- **File-Drawer Problem:** Non-significant results are never published (end up in the file drawer)
- **Underpowered Studies:** Sample sizes are often too small → many pilots with small samples give you more significant results (economics of science), but it is much more likely that these results are not valid

Open Science Practices in the Research Cycle

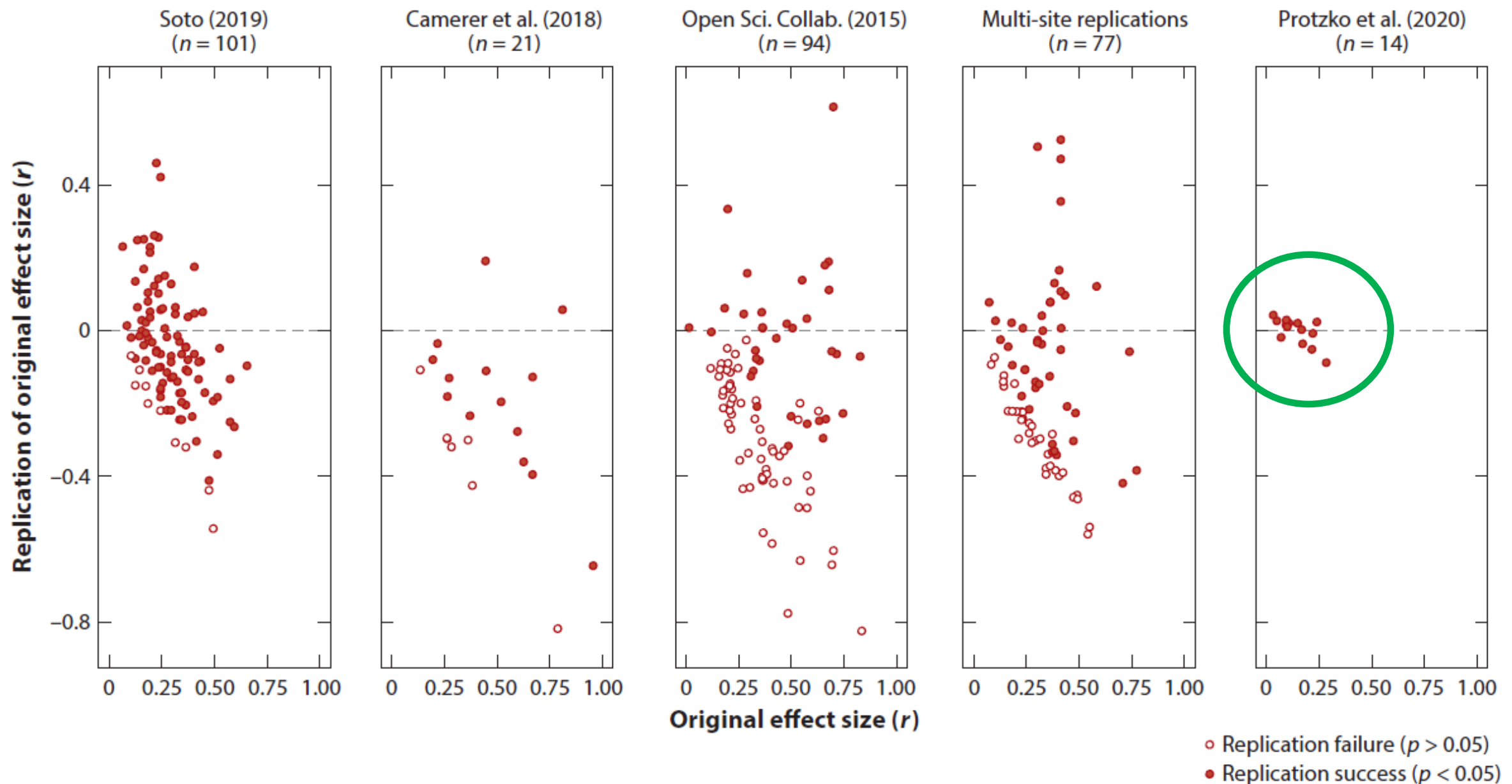


- **p-Hacking:** Doing everything to get that damn p-value
- **HARKing:** Hypthesizing After Results are Known – quite common distortion of assumptions held a priori
- **File-Drawer Problem:** Non-significant results are never published (end up in the file drawer)
- **Underpowered Studies:** Sample sizes are often too small → many pilots with small samples give you more significant results (economics of science), but it is much more likely that these results are not valid

Questionable
Research
Practices

How much replicability can we achieve?

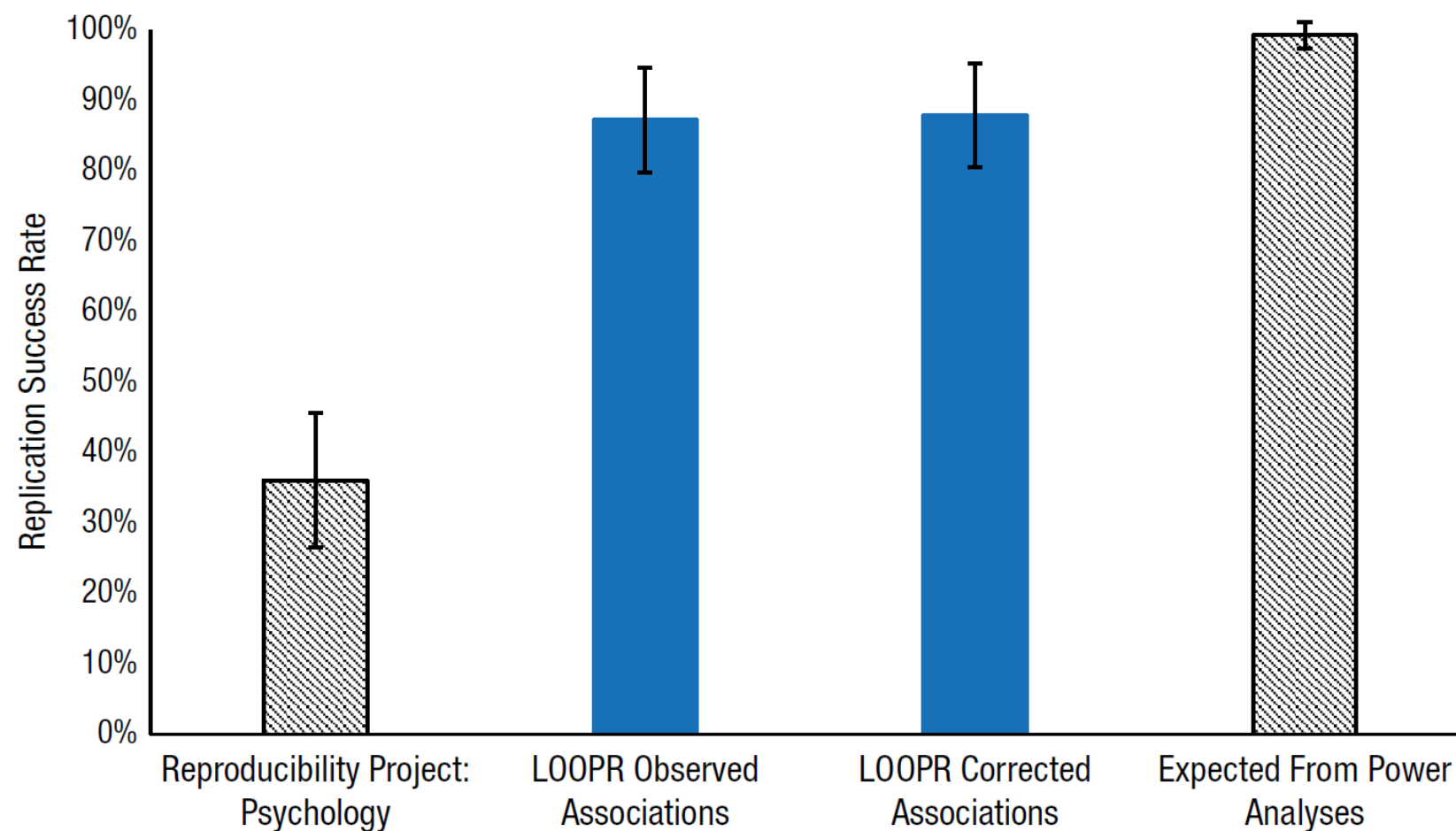
Success rate with stringent methodology: **90% replicability, effect sizes at 97% of original**



Replication Successes in Some (Sub-)Fields

Life Outcomes of Personality Replication Project (LOOPR):

87% of successful replications, effects sizes at 77% of the original studies



(Other) Ways to Move Forward

- **Top-down action:** Institutions, journals and funding bodies should change their guidelines to improve research practices: quality vs. quantity, publication options, funding requirements.
- **Field-specific standards** for good scientific practice should be adjusted to include open-science/transparent practices where appropriate.
- **Offering attractive solutions:** *Sequential analyses* allow to achieve reasonable sample sizes in an efficient manner. With proper documentation, results can be statistically checked at pre-defined sampling stages.
 - **Traditional statistics:**
Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7), 701–710. <https://doi.org/10.1002/ejsp.2023>
 - **Bayesian statistics:**
Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1), 128–142. <https://doi.org/10.3758/s13423-017-1230-y>

Summary

- **Questionable research practices** threaten the validity of research records in different fields.
- **Replication projects** demonstrate that in some fields there are substantial problems with the replicability of results.
- **Open science practices** can help improve the reliability of the research record. Practices have to be realistic and easily actionable. Focusing on benefits can speed up the transition.
- **Incentive structures and guidelines** have to change to implement practices in the long run.

Trouble for Science

MANY COMMERCIAL ANTIBODY-BASED
IMMUNOASSAYS ARE UNRELIABLE

PROBLEMS WITH THE p -VALUE AS AN
INDICATOR OF SIGNIFICANCE

OVERFEEDING OF LABORATORY RODENTS
COMPROMISES ANIMAL MODELS

REPLICATION STUDY FAILS TO REPRODUCE
MANY PUBLISHED RESULTS

CONTROLLED TRIALS SHOW BUNSEN
BURNERS MAKE THINGS COLDER

Two additional cartoons
for self-study

<https://xkcd.com/882/>

<https://xkcd.com/808/>

<https://xkcd.com/1574/>

Creative Commons Attribution-
NonCommercial 2.5 Generic
(CC BY-NC 2.5)