SCIENTIFIC ECOSYSTEMS AND RESEARCH REPRODUCIBILITY

Marcus Munafò
A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic

Aaron Mobley¹, Suzanne K. Linder², Russell Braeuer¹, Lee M. Ellis¹,³*, Leonard Zwelling⁴*

¹Department of Cancer Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America. ²Department of General Internal Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America. ³Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America. ⁴Department of Experimental Therapeutics, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America

CORRESPONDENCE

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science

Open Science Collaboration¹

Abstract

Reproducibility is a defining feature of science. However, because of strong incentives for innovation and weak incentives for confirmation, direct replication is rarely practiced or published. The Reproducibility Project is an open, large-scale, collaborative effort to systematically examine the rate and predictors of reproducibility in psychological science. So far, 72 volunteer researchers from 41 institutions have organized to openly and transparently replicate studies published in three prominent psychological journals in 2008. Multiple methods will be used to evaluate the findings, calculate an empirical rate of replication, and investigate factors that predict reproducibility. Whatever the result, a better understanding of reproducibility will ultimately improve confidence in scientific methodology and findings.
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.
Prediction market on the outcomes of the Reproducibility Project: Psychology

Successful replications are shown in black, unsuccessful replications in red.

Dreber et al. (2015). PNAS, 112, 1534
“Scientists may be in the business of laughing at their predecessors, but owing to an array of human mental dispositions, few realize that someone will laugh at their beliefs in the (disappointingly near) future”

Incentive Structures

We Knew the Future All Along: Scientific Hypothesizing is Much More Accurate Than Other Forms of Precognition—A Satire in One Part

Arina K. Bones
University of Darache, Monte Carlo, Monaco

Scientists behaving badly

To protect the integrity of science, we must look beyond falsification, fabrication and plagiarism, to a wider range of questionable research practices, argue Brian C. Martinson, Melissa S. Anderson and Raymond de Vries.

“Certain features of the working environment of science may have unexpected and potentially detrimental effects on the ethical dimensions of scientists’ work”

(i) Limbo
(ii) Overselling
(iii) Post-hoc storytelling
(iv) P-value fishing
(v) Creative outliers
(vi) Plagiarism
(viii) Non-publication
(viii) Partial publication
(ix) Falsification

http://blogs.discovermagazine.com/neuroskeptic/2013/10/16/the-f-problem
Using the same method as in Study 1, we asked 20 34 University of Pennsylvania undergraduates to listen only to either “When I’m Sixty-Four” by The Beatles or “Kalimba” or “Hot Potato” by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. Then, in an ostensibly unrelated task, they indicated only their birth date (mm/dd/yyyy) and how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with “computers are complicated machines,” their father’s age, their mother’s age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as “the good old days,” and their gender. We used father’s age to control for variation in baseline age across participants.

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

“…nearly as many unique analysis pipelines as there were studies in the sample…”

Carp (2012). Neuroimage, 63, 289-300.
Incentive Structures

ANALYSIS

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button1,2, John P. A. Ioannidis3, Claire Mokrysz1, Brian A. Nosek4, Jonathan Flint5, Emma S. J. Robinson6 and Marcus R. Munafò7

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.

Studies from top-ranked UK institutions perform worse on reporting of measures to reduce the risk of bias than studies selected at random from PubMed…

# US studies may overestimate effect sizes in softer research

Daniele Fanelli\textsuperscript{a,1} and John P. A. Ioannidis\textsuperscript{b,c,d}

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Nonbehavioral ($k = 40$, $n = 566$)</th>
<th>Behavioral, all ($k = 42$, $n = 608$)</th>
<th>Biobehavioral ($k = 20$, $n = 308$)</th>
<th>Behavioral ($k = 22$, $n = 300$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.42 [0.40, 0.46]</td>
<td>0.55 [0.51, 0.56]</td>
<td>0.51 [0.47, 0.54]</td>
<td>0.57 [0.50, 0.59]</td>
</tr>
<tr>
<td>United States vs.</td>
<td>-0.02 [-0.06, 0.00]</td>
<td>0.03 [0.02, 0.06]</td>
<td>0.03 [0.00, 0.07]</td>
<td>0.04 [0.01, 0.07]</td>
</tr>
<tr>
<td>rest</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study size (SE)</td>
<td>0.43 [0.27, 0.53]</td>
<td>0.11 [0.07, 0.23]</td>
<td>0.20 [0.11, 0.31]</td>
<td>0.06 [0.01, 0.29]</td>
</tr>
<tr>
<td>Pub. order</td>
<td>0.02 [0.00, 0.03]</td>
<td>0.00 [-0.01, 0.01]</td>
<td>0.01 [0.00, 0.05]</td>
<td>0.00 [-0.02, 0.01]</td>
</tr>
<tr>
<td>USA*SE</td>
<td>-0.21 [-0.47, 0.22]</td>
<td>-0.19 [-0.31, -0.03]</td>
<td>-0.16 [-0.34, 0.12]</td>
<td>-0.22 [-0.46, -0.02]</td>
</tr>
<tr>
<td>USA*pub. order</td>
<td>-0.02 [-0.05, 0.01]</td>
<td>0.00 [-0.02, 0.03]</td>
<td>-0.02 [-0.06, 0.01]</td>
<td>0.01 [-0.02, 0.05]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature, Science</td>
<td>$26,212</td>
<td>$26,006</td>
<td>$25,781</td>
<td>$25,365</td>
<td>$33,990</td>
<td>$36,658</td>
<td>$38,908</td>
<td>$43,783</td>
<td>$43,783</td>
</tr>
<tr>
<td>PLOS One</td>
<td>$1,096</td>
<td>$1,086</td>
<td>$1,035</td>
<td>$994</td>
<td>$991</td>
<td>$915</td>
<td>$941</td>
<td>$984</td>
<td>$984</td>
</tr>
<tr>
<td>MIS Quarterly</td>
<td>$2,613</td>
<td>$2,570</td>
<td>$2,553</td>
<td>$2,654</td>
<td>$2,876</td>
<td>$2,861</td>
<td>$2,992</td>
<td>$2,938</td>
<td>$2,938</td>
</tr>
<tr>
<td>JASIST</td>
<td>$1,737</td>
<td>$1,758</td>
<td>$1,741</td>
<td>$1,887</td>
<td>$2,066</td>
<td>$2,303</td>
<td>$2,435</td>
<td>$2,488</td>
<td>$2,488</td>
</tr>
<tr>
<td>Journal of Documentation</td>
<td>$1,082</td>
<td>$1,087</td>
<td>$1,042</td>
<td>$1,111</td>
<td>$1,167</td>
<td>$1,265</td>
<td>$1,329</td>
<td>$1,408</td>
<td>$1,408</td>
</tr>
<tr>
<td>Library Hi Tech</td>
<td>$781</td>
<td>$775</td>
<td>$726</td>
<td>$741</td>
<td>$740</td>
<td>$768</td>
<td>$795</td>
<td>$783</td>
<td>$783</td>
</tr>
<tr>
<td>LIBRI</td>
<td>$650</td>
<td>$644</td>
<td>$577</td>
<td>$560</td>
<td>$538</td>
<td>$509</td>
<td>$517</td>
<td>$484</td>
<td>$484</td>
</tr>
</tbody>
</table>

* All the amounts are full amount (in USD) awarded to the first author
Why Science Is Not Necessarily Self-Correcting

John P. A. Ioannidis
Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy, Stanford University School of Medicine, and Department of Statistics, Stanford University School of Humanities and Sciences

“Among 83 articles recommending effective interventions, 40 had not been subject to any attempt at replication…”

Primary study authors of significant studies are more likely to believe that a strong association exists in a heterogeneous meta-analysis compared with methodologists.

Orestis A. Panagiotou\textsuperscript{a}, John P.A. Ioannidis\textsuperscript{b,c,d,e,*}
How citation distortions create unfounded authority: analysis of a citation network

Steven A Greenberg, associate professor of neurology

Investigated citation network of papers addressing the belief that B amyloid, a protein accumulated in the brain in Alzheimer’s disease, is produced by and injures skeletal muscle of patients with inclusion body myositis.

Abstracts often “spin” results to give impression that results are positive when they are not.

Citation inflation exists for both “positive” studies and “claim” studies in this literature.

True both within this literature (A, B) and in the wider (Web of Science) literature (C, D).

Two positive trials, four neutral trials, two negative trials (stopped early for safety concerns).
Real Scientific Method

The Scientific Method

Observe natural phenomena → Formulate Hypothesis → Modify Hypothesis → Test hypothesis via rigorous Experiment → Establish Theory based on repeated validation of results

The Actual Method

Make up Theory based on what Funding Agency Manager wants to be true → Design minimum experiments that will **prove show?** suggest Theory is true → Modify Theory to fit data → Publish Paper: rename Theory a "Hypothesis" and pretend you used the Scientific Method → Defend Theory despite all evidence to the contrary

www.phdcomics.com
Scientific rigor and the art of motorcycle maintenance

Marcus Munafò, Simon Noble, William J Browne, Dani Brunner, Katherine Button, Joaquim Ferreira, Peter Holmans, Douglas Langbehn, Glyn Lewis, Martin Lindquist, Kate Tilling, Eric-Jan Wagenmakers & Robi Blumenstein

The reliability of scientific research is under scrutiny. A recently convened working group proposes cultural adjustments to incentivize better research practices.

Munafò et al. (2014), Nat Biotech, 32, 871-873.
Open Science

- Open Data
- Open Source
- Open Methodology
- Open Peer Review
- Open Access
- Open Educational Resources
In 2000 the National Heart Lung, and Blood Institute required the registration of primary outcome on ClinicalTrials.gov for all their grant-funded activity.
Introduction of badges for open practices at *Psychological Science* followed by a steep increased in data sharing.

A manifesto for reproducible science

Marcus R. Munafò¹,²*, Brian A. Nosek³,⁴, Dorothy V. M. Bishop⁵, Katherine S. Button⁶, Christopher D. Chambers⁷, Nathalie Percie du Sert⁸, Uri Simonsohn⁹, Eric-Jan Wagenmakers¹⁰, Jennifer J. Ware¹¹ and John P. A. Ioannidis¹²,¹³,¹⁴

Table 1 | A manifesto for reproducible science.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Proposal</th>
<th>Examples of initiatives/potential solutions (extent of current adoption)</th>
<th>Stakeholder(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>Protecting against cognitive biases</td>
<td>All of the initiatives listed below (* to ****)</td>
<td>J, F</td>
</tr>
<tr>
<td></td>
<td>Blinding (***)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Improving methodological training</td>
<td>Rigorous training in statistics and research methods for future researchers (*)</td>
<td>I, F</td>
</tr>
<tr>
<td></td>
<td>Rigorous continuing education in statistics and methods for researchers (•*)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Independent methodological support</td>
<td>Involvement of methodologists in research (***)</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>Independent oversight (*)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Collaboration and team science</td>
<td>Multi-site studies/distributed data collection (•)</td>
<td>I, F</td>
</tr>
<tr>
<td></td>
<td>Team-science consortia (•)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reporting and dissemination</td>
<td>Promoting study pre-registration</td>
<td>Registered Reports (•)</td>
<td>J, F</td>
</tr>
<tr>
<td></td>
<td>Open Science Framework (•)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Improving the quality of reporting</td>
<td>Use of reporting checklists (•)</td>
<td>J</td>
</tr>
<tr>
<td></td>
<td>Protocol checklists (•)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Protecting against conflicts of interest</td>
<td>Disclosure of conflicts of interest (***), Exclusion/containment of financial and non-financial conflicts of interest (•)</td>
<td>J, F,R</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>Encouraging transparency and open science</td>
<td>Open data, materials, software and so on (* to ***)</td>
<td>J, F, R</td>
</tr>
<tr>
<td></td>
<td>Pre-registration (***) for clinical trials, * for other studies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation</td>
<td>Diversifying peer review</td>
<td>Preprints (* in biomedical/behavioural sciences, **** in physical sciences)</td>
<td>J</td>
</tr>
<tr>
<td></td>
<td>Pre- and post-publication peer review, for example, PubMed, PubMed Commons (•)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incentives</td>
<td>Rewarding open and reproducible practices</td>
<td>Badges (*), Registered Reports (•), Transparency and Openness Promotion guidelines (•), Funding replication studies (•), Open science practices in hiring and promotion (•)</td>
<td>J, F, L</td>
</tr>
</tbody>
</table>

Estimated extent of current adoption: *, <5%; **, 5-20%; ***, 20-40%; ****, >40%. Abbreviations for key stakeholders: J, journals; P, publishers; F, funders; L, institutions; R, regulators.

Munafò et al. (2017). Nat Hum Behav, 1, 0021.
UK Reproducibility Network

Understand factors that contribute to poor research reproducibility
Provide training and disseminate best practice
Support and test interventions to improve reproducibility
Ensure coordination with stakeholders

• Launched March 2019
• Local network leads at >40 UK institutions
• Supported by a range of stakeholders
Acknowledgements

marcus.munafo@bristol.ac.uk
@MarcusMunafo
@BristolTARG
@UKRepro

http://www.bristol.ac.uk/expsych/research/brain/targ
https://www.bristol.ac.uk/psychology/research/ukrn/

Tobacco and Alcohol Research Group:

Angela Attwood Senior Lecturer
Anna Blackwell Postdoc
Claire Braboszcz Postdoc
Laura Brocklebank Postdoc
Sarah Burrows-Weeks Administrator
Katie De-Loyde Research Assistant
Katie Drax PhD Student
Maddy Dyer PhD Student
Kayleigh Easey PhD Student
Andy Eastwood PhD Student
Jenn Ferrar Postdoc
Will Gawned Administrator
Elis Haan PhD Student
Abigail Jackson Postdoc
Andy Keavey PhD Student
Jasmine Khouja PhD Student
Rebecca Lawn PhD Student
Liam Mahedy Postdoc
Osama Mahmoud Postdoc
Joe Matthews Research Assistant
Olivia Maynard Lecturer
Osama Mahmoud PhD Student
Vicky Rice PhD Student
Laura Schellhas Postdoc
Hannah Sallis PhD Student
Carlos Sillero Postdoc
Andy Skinner Research Assistant
Caroline Skirrow Postdoc
Chris Stone Postdoc
Steph Suddell Research Assistant
Jackie Thompson Postdoc
Sophie Turnbull Postdoc
Robyn Wootton Postdoc