

**„Always good to start  
with a joke“**

---

**Meike Steinhilber**



**UM... PSYCHOTHERAPIST  
IS ONE WORD.**



# Replication Crisis – Psychology

---

**Seminar: How do I lie with statistics?**

**Supervisor: Prof. Dr. Ullrich Köthe**

**Meike Steinhilber**

**21. November 2019**

# Structure

---

- I. The Crisis
- II. Is there really a Crisis?
- III. Consequences & Changes

**CRISIS!**  
**WHAT CRISIS?**



# I. The Crisis

---

**(1) Introduction**

**(2) Emergence and Development of the Crisis**

# I. The Crisis

published research findings false



Ungefähr 1.460.000 Ergebnisse (0,07 Sek.)

## [HTML] Why most published research findings are false

[JPA Ioannidis](#) - PLoS medicine, 2005 - journals.plos.org

There is increasing concern that most current **published research findings** are **false**. The probability that a **research** claim is true may depend on **study** power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships ...

☆ Zitiert von: 7698 Ähnliche Artikel Alle 154 Versionen

## [HTML] Most published research findings are false—but a little replication goes a long way

[R Moonesinghe](#), [MJ Khoury](#), [ACJW Janssens](#) - PLoS medicine, 2007 - journals.plos.org

This is an open-access article distributed under the terms of the Creative Commons Public Domain declaration which stipulates that, once placed in the public domain, this work may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any ...

☆ Zitiert von: 231 Ähnliche Artikel Alle 19 Versionen

## Are most published research findings false?

[A Diekmann](#) - Jahrbücher für nationalökonomie und statistik, 2011 - degruyter.com

In a provocative article Ioannidis (2005) argues that, in disciplines employing statistical tests of significance, professional journals report more **wrong** than true significant **results**. This short note sketches the argument and explores under what conditions the assertion holds ...

☆ Zitiert von: 17 Ähnliche Artikel Alle 12 Versionen

## [HTML] Why most published research findings are false: problems in the analysis

[S Goodman](#), [S Greenland](#) - PLoS medicine, 2007 - journals.plos.org

It is only about 25 years since the abolition or abandonment of TB sanatoria in South Africa. It is likely that there are people in South Africa who are familiar with the idea of isolation of TB patients from the general public emanating from the era of TB sanatoria, and it is ...

☆ Zitiert von: 74 Ähnliche Artikel Alle 7 Versionen

## Why most published research findings are false: author's reply to Goodman and Greenland

[JPA Ioannidis](#) - PLoS medicine, 2007 - journals.plos.org

We applaud PLoS editors for their commitment to **publishing** high-quality systematic reviews (SRs)[1]. Moher et al.[2] clearly documented the inconsistent quality of reporting of SRs. With more than 2,500 SRs **published** every year, low-quality or outdated reviews may mislead ...

☆ Zitiert von: 72 Ähnliche Artikel Alle 6 Versionen

# The Research Reproducibility Crisis and Economics of Science

Zacharias Maniadis, Fabio Tufano

*The Economic Journal*, Volume 127, Issue 605, October 2017, Pages F200–F208,

<https://doi.org/10.1111/eoj.12526>

**Published:** 24 October 2017

“ Cite   🔑 Permissions   ➦ Share ▼

### Abstract

To address the increasing concern about research reproducibility, cross-fertilisation across economics and other disciplines is likely to have far-reaching benefits. Our brief summary focuses on two areas in which a mutual investment in investigating possible cross-disciplinary synergies could benefit the scientific endeavour as a whole. First, the discipline of economic design has much to contribute to the discussion of possible reforms in science. Second, the empirical methodology of meta-research can inform practices to assess the validity of the economics literature.

# Artificial intelligence faces reproducibility crisis

Matthew Hutson

+ See all authors and affiliations

Science 16 Feb 2018:

Vol. 359, Issue 6377, pp. 725-726

DOI: [10.1126/science.359.6377.725](https://doi.org/10.1126/science.359.6377.725)

Article

Figures & Data

Info & Metrics

eLetters

 PDF

## Summary

The booming field of artificial intelligence (AI) is grappling with a replication crisis, much like the ones that have afflicted psychology, medicine, and other fields over the past decade. Just because algorithms are based on code doesn't mean experiments are easily replicated. Far from it.

Unpublished codes and a sensitivity to training conditions have made it difficult for AI researchers to reproduce many key results. That is leading to a new conscientiousness about research methods and publication protocols. Last week, at a meeting of the Association for the Advancement of Artificial Intelligence in New Orleans, Louisiana, reproducibility was on the agenda, with some teams diagnosing the problem—and one laying out tools to mitigate it.

[View Full Text](#)



*Annual Review of Chemical and Biomolecular Engineering*

# Does Chemical Engineering Research Have a Reproducibility Problem?

Rebecca Han, Krista S. Walton, and David S. Sholl

School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0100, USA; email: david.sholl@chbe.gatech.edu

## The Replication Crisis and Chemistry Education Research

Melanie M. Cooper\*

Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States

*“Why, you may ask, am I discussing these social psychology studies in a chemistry education editorial? The answer lies in the fact that many of the methods used in these studies are also found in papers published in this Journal.”*

# Raise standards for preclinical cancer research

C. Glenn Begley & Lee M. Ellis 

*Nature* **483**, 531–533(2012) | [Cite this article](#)

**27k** Accesses | **1285** Citations | **2014** Altmetric | [Metrics](#)

 A Clarification to this article was published on 02 May 2012

*“[...] Scientific findings were confirmed in only 6 (11%) cases. Even knowing the limitations of preclinical research, this was a shocking result.”*

(53 studies have been tried to replicate)

*“A team at Bayer HealthCare in Germany last year reported that only about 25% of published preclinical studies could be validated to the point at which projects could continue.”*

Crosstalk

# Never Waste a Good Crisis: Confronting Reproducibility in Translational Research

Daniel J. Drucker<sup>1</sup>  

 **Show more**

<https://doi.org/10.1016/j.cmet.2016.08.006>

[Get rights and content](#)

Under an [Elsevier user license](#)

[open archive](#)

---

The lack of reproducibility of preclinical [experimentation](#) has implications for sustaining trust in and ensuring the viability and funding of the academic research enterprise. Here I identify problematic behaviors and practices and suggest solutions to enhance reproducibility in [translational research](#).

# Replication Crisis – Why focus on Psychology?

- My own subject is psychology
- Crisis is mostly discussed in the field of psychology

The screenshot shows a search engine interface with the query 'replication crisis' entered in the search bar. Below the search bar is a dropdown menu with several suggestions: 'replication crisis psychology', 'replication crisis publication', 'replication crisis nosek', 'replication crisis social psychology', 'replication crisis psychology suffering', 'knowledge construction replication crisis', 'original studies replication crisis', and 'ego depletion replication crisis'. Below the suggestions are three search results, each starting with a green checkmark icon. The first result is titled 'Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?' and is by SE Maxwell, MY Lau, and GS Howard, published in American Psychologist in 2015. The second result is titled 'Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis' and is by PE Shrout and JL Rodgers, published in Annual review of psychology in 2018. The third result is titled 'Psychology's replication crisis and the grant culture: Righting the ship' and is by SO Lilienfeld, published in Perspectives on psychological science in 2017. Each result includes a brief abstract and metadata such as citation counts and version numbers.

replication crisis

- replication crisis **psychology**
- replication crisis **publication**
- replication crisis **nosek**
- replication crisis **social psychology**
- replication crisis **psychology suffering**
- knowledge construction** replication crisis
- original studies** replication crisis
- ego depletion** replication crisis

✓ **Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?**  
[SE Maxwell](#), [MY Lau](#), [GS Howard](#) - American Psychologist, 2015 - psycnet.apa.org  
Psychology has recently been viewed as facing a **replication crisis** because efforts to **replicate** past study findings frequently do not show the same result. Often, the first study showed a statistically significant result but the **replication** does not. Questions then arise ...  
☆ 99 Zitiert von: 362 Ähnliche Artikel Alle 6 Versionen

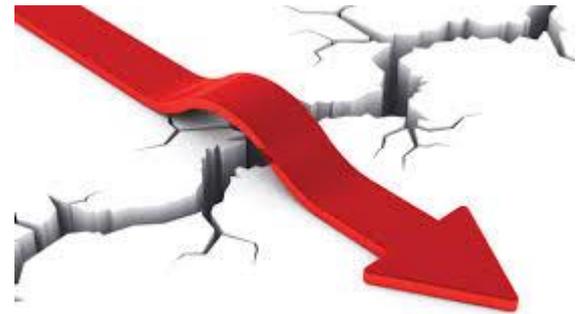
✓ **Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis**  
[PE Shrout](#), [JL Rodgers](#) - Annual review of psychology, 2018 - annualreviews.org  
Psychology advances knowledge by testing statistical hypotheses using empirical observations and data. The expectation is that most statistically significant findings can be **replicated** in new data and in new laboratories, but in practice many findings have **replicated** ...  
☆ 99 Zitiert von: 96 Ähnliche Artikel Alle 11 Versionen

✓ **Psychology's replication crisis and the grant culture: Righting the ship**  
[SO Lilienfeld](#) - Perspectives on psychological science, 2017 - journals.sagepub.com  
The past several years have been a time for soul searching in psychology, as we have gradually come to grips with the reality that some of our cherished findings are less robust than we had assumed. Nevertheless, the **replication crisis** highlights the operation of ...  
☆ 99 Zitiert von: 55 Ähnliche Artikel Alle 6 Versionen

# Emergence and Development of the Crisis

---

- **2005**
  - Beginning of the crisis in general
- **2011**
  - Beginning of the crisis in psychology
- **2018**
  - *“Psychology’s Renaissance”*



[1,2]

# Beginning of the Crisis in 2005

---

**John Ioannidis, PLoS Medicine**

## **“Why Most Published Research Findings Are False”**

*“[A] research finding is less likely to be true when the **studies conducted in a field are smaller**; when **effect sizes are smaller**; when there is a greater number and lesser preselection of **tested relationships**; where there is **greater flexibility** in designs, definitions, outcomes, and analytical modes; when there is **greater financial and other interest and prejudice**; and when more teams are involved in a scientific field in **chase of statistical significance.**”*

[2]

# Beginning of the Crisis in Psychology in 2011

---

- **The case of Daryl Bem**
- **The case of Diederik Stapel**
- **Paper “False-Positive Psychology” (Simmons et al. 2011)**
- **Brian Nosek**
  - Developed the online platform “Open Science Framework”
  - Organized a large replication study

[1]

# The Case of Daryl Bem

---



- **The Journal of Personality and Social Psychology published a paper by Daryl Bem**
  - with a transparently outlandish claim, that people can be influenced by an unforeseeable future event
  - Bem found that participants were better able to recall words that they were later randomly assigned to rehearse
    - Obviously, a replication was not possible ;-)

[1]



# The Case of Diederik Stapel

---

- **Former professor in social psychology in the Netherlands**
- **He was one of the most prominent researchers in social psychology**
- **Confessed to decades of data fabrication**
  - **Co-workers got suspicious, because they never saw any test persons in the labs**



[1]

# Paper “False-Positive Psychology”

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

Psychological Science  
22(11) 1359–1366  
© The Author(s) 2011  
Reprints and permission:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/0956797611417632  
<http://pss.sagepub.com>  


Joseph P. Simmons<sup>1</sup>, Leif D. Nelson<sup>2</sup>, and Uri Simonsohn<sup>1</sup>

<sup>1</sup>The Wharton School, University of Pennsylvania, and <sup>2</sup>Haas School of Business, University of California, Berkeley

### Abstract

In this article, we accomplish two things. First, we show that despite empirical psychologists’ nominal endorsement of a low rate of false-positive findings ( $\leq .05$ ), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. Second, we suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. The solution involves six concrete requirements for authors and four guidelines for reviewers, all of which impose a minimal burden on the publication process.

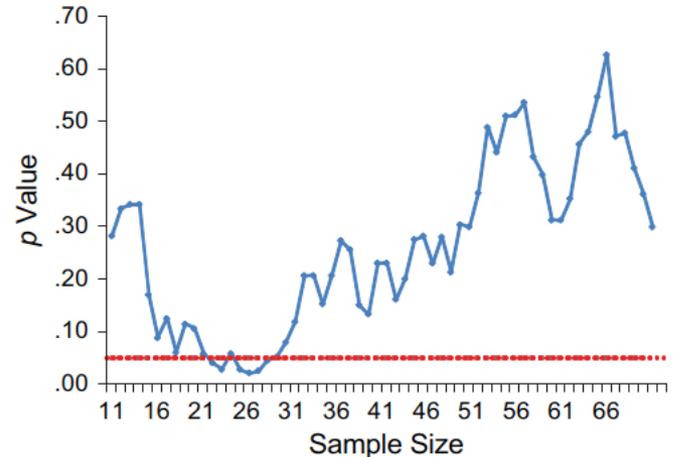
[3]

## “How Bad Can It Be?”

# “How Bad Can It Be?”

---

- **Researchers degrees of freedom**
- **Flexibility in choosing**
  - among dependent variables
  - sample size
    - *stop of data collection*
    - *outlier analysis*
  - covariates
  - and reporting subsets of experimental conditions
  - ...



[3]

## I. The Crisis

**Table I.** Likelihood of Obtaining a False-Positive Result

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
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%

Note: The table reports the percentage of 15,000 simulated samples in which at least one of a set of analyses was significant. Observations were drawn independently from a normal distribution. Baseline is a two-condition design with 20 observations per cell. Results for Situation A were obtained by conducting three  $t$  tests, one on each of two dependent variables and a third on the average of these two variables. Results for Situation B were obtained by conducting one  $t$  test after collecting 20 observations per cell and another after collecting an additional 10 observations per cell. Results for Situation C were obtained by conducting a  $t$  test, an analysis of covariance with a gender main effect, and an analysis of covariance with a gender interaction (each observation was assigned a 50% probability of being female). We report a significant effect if the effect of condition was significant in any of these analyses or if the Gender  $\times$  Condition interaction was significant. Results for Situation D were obtained by conducting  $t$  tests for each of the three possible pairings of conditions and an ordinary least squares regression for the linear trend of all three conditions (coding: low =  $-1$ , medium =  $0$ , high =  $1$ ).

[3]

# Published Abstract & Reality

---

**Table 3.** Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

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 ( $M_s = 20.3$  and  $21.2$ , respectively),  $F(1, 18) = 1.01, p = .33$ .

---

# Brian Nosek



#### Search and Discover

Find papers, data, and materials to inspire your next research project. Search public projects to build on the work of others and find new collaborators.



#### Design Your Study

Start a project and add collaborators, giving them access to protocols and other research materials. Built-in version control tracks the evolution of your study.



#### Collect and Analyze Data

Store data, code, and other materials in OSF Storage, or connect your Dropbox or other third-party account. Every file gets a unique, persistent URL for citing and sharing.



#### Publish Your Reports

Share papers in OSF Preprints or a community-based preprint provider, so others can find and cite your work. Track impact with metrics like downloads and view counts.

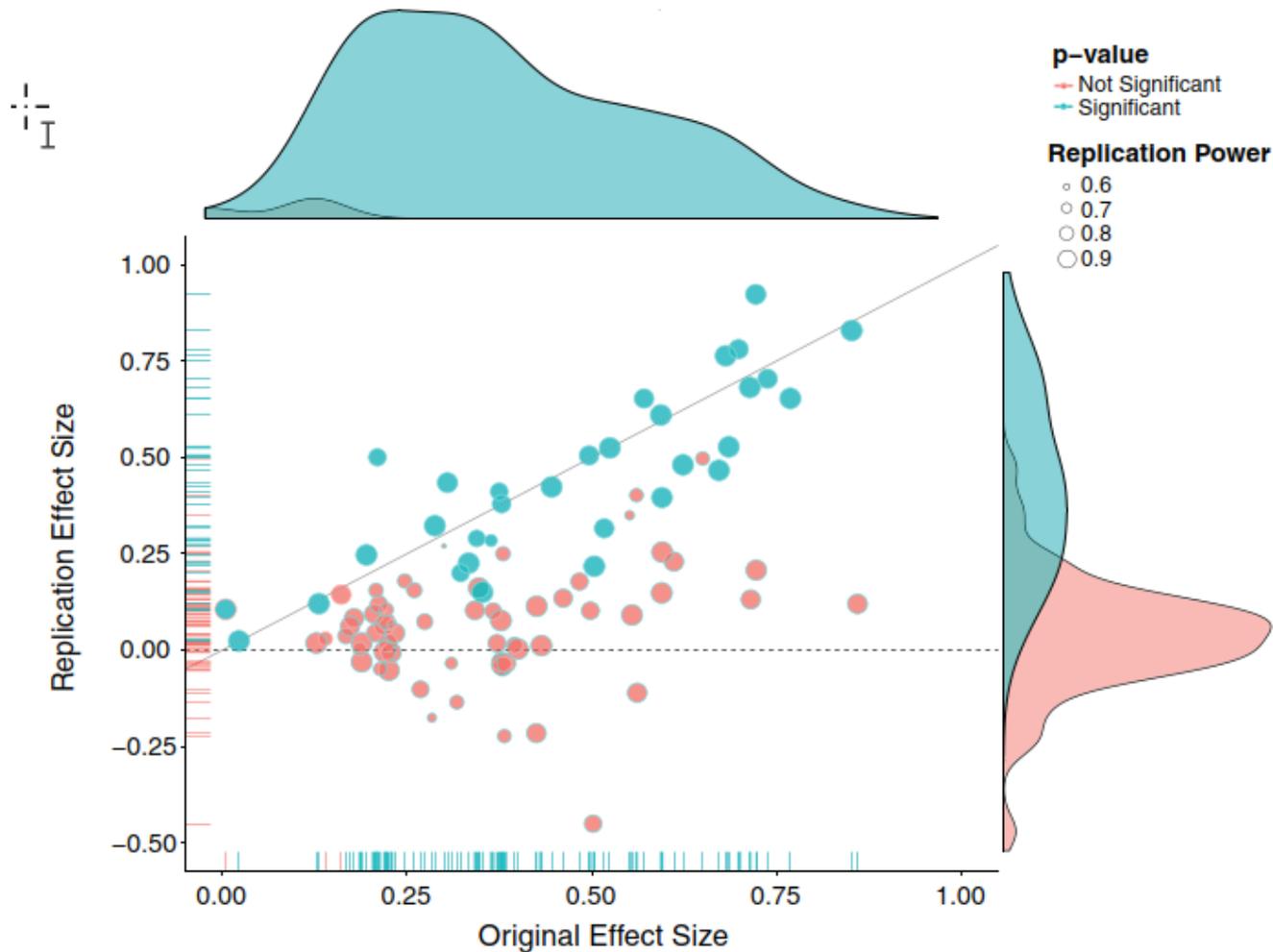
- **Developed an online platform**
  - Open Science Framework (OSF)
  - released in 2012
  - allowed researchers to more transparently record, share and report their work
- **Organized replication study (Open Sci. Collab., 2012, 2015)**

# Replication Study (Open Sci. Collab., 2015)

---

- **270 contributors**
- **100 studies were attempted to replicate**
  - random selection from 3 high profiled journals from 2008
  - experimental & correlational studies
- **Obtaining original material, if available**
- **Increased power of the original design**
- **Internal review**

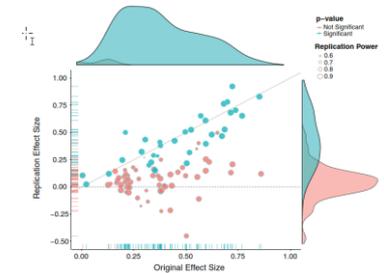
## I. The Crisis



**Fig. 3. Original study effect size versus replication effect size (correlation coefficients).** Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

[4]

# Replication Study (Open Sci. Collab., 2015)



- **Replication effects were half the magnitude of original effects**
  - substantial decline
- **97% of original studies had statistically significant results**
  - but only 36 % of the replications had statistically significant results
- **47% of original effect sizes were in the 95% confidence interval of the replication effect size**
- **“If no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects.”**
  - **Are at least 1/3 of results in psychology false alarms?**

[4]

# Quick Overview

---

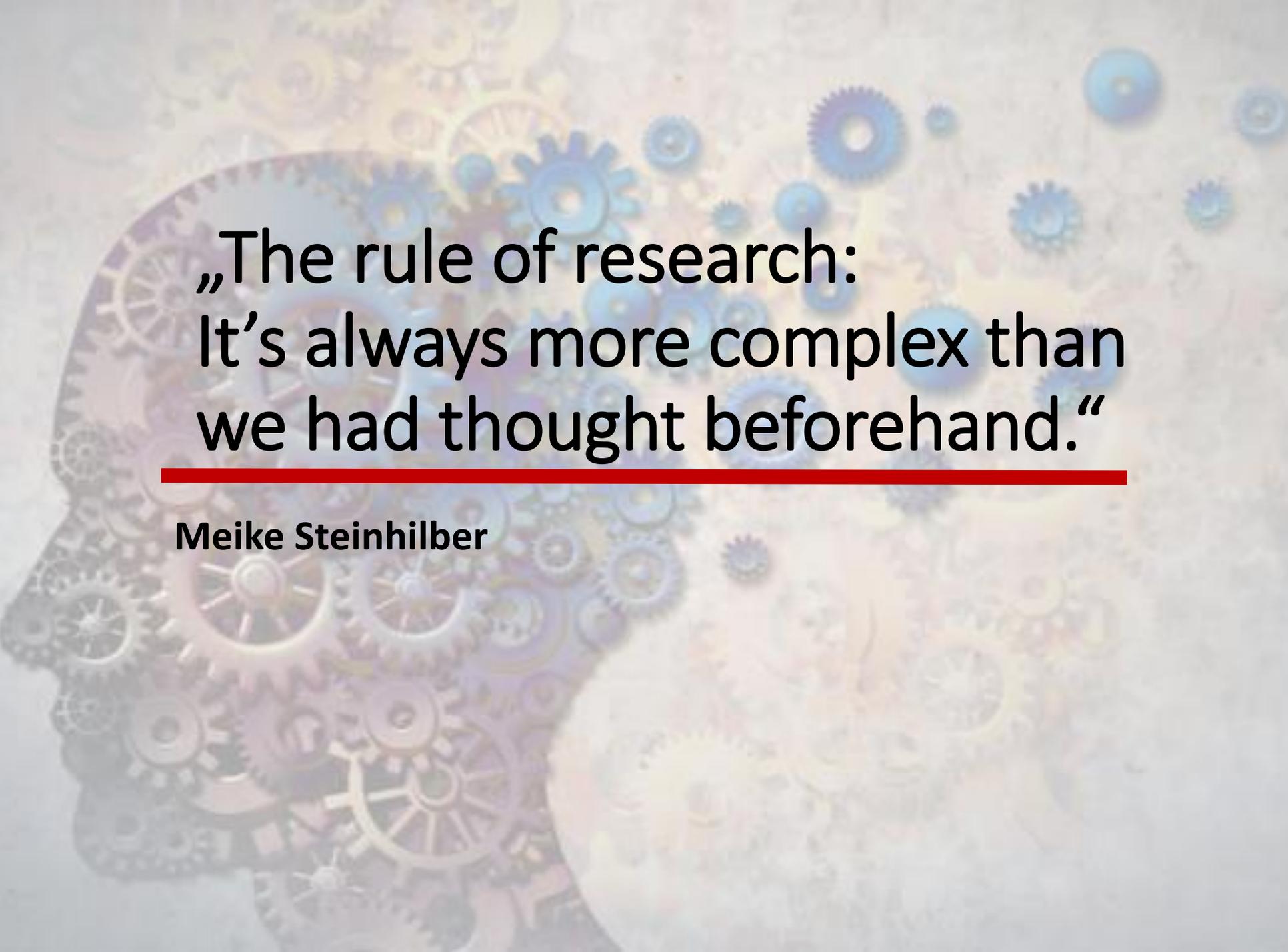
- **The case of Daryl Bem**
- **The case of Diederik Stapel**
  - a rotten apple spoils the bunch
- **Paper “False-Positive Psychology”**
  - manipulation of results is easier, than many would have thought
- **Replication Study (Open Sci. Collab., 2015)**
  - symptoms of actual systematic problems

Level  
of  
Proof



**CRISIS!**

[1]



**„The rule of research:  
It's always more complex than  
we had thought beforehand.“**

---

**Meike Steinhilber**

# WHAT CRISIS?

*“In contrast to the prevalent sentiment, we will argue that the claim of a replicability crisis is **greatly exaggerated** and that the hope that such a crisis (**if it ever existed**) could be solved by increasing the number of exact replications is misplaced.”*

[5]

*“Indeed, the data are consistent with the opposite conclusion, namely, that the **reproducibility of psychological science is quite high.**”*

[6]

# Structure

---

- I. The Crisis
- II. Is there really a Crisis?
- III. Consequences & Changes

**CRISIS!**  
**WHAT CRISIS?**



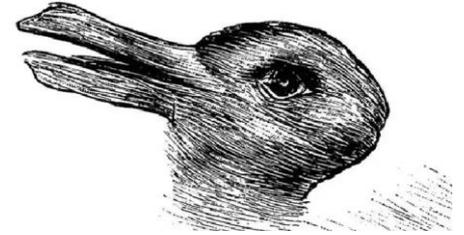
## II. Is there really a Crisis?

---

- (1) Illusion of an exact replication**
- (2) Critique on the replication study (Open Sci. Collab., 2015)**
- (3) What is the actual problem, if not replication?**

# Illusion of an exact Replication

---



- **Forms of replication studies**

- exact replication
- conceptual replication

- **Exact replication**

- = Repetition of an experimental procedure to a degree **as exact as possible**.
- Change of test persons & researcher team

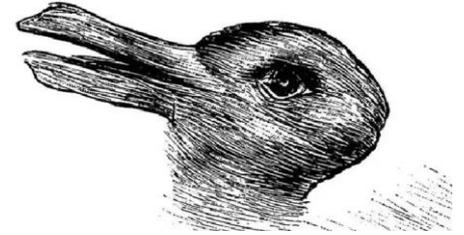
- **Conceptual replication**

- = Re-test the same theoretical idea or hypothesis repeatedly
- but use different populations, different ways of manipulating variables, different ways of measuring variables, or using different study designs

[5, 6, 7]

# Illusion of an exact Replication

---



- **Conceptual replication**
  - positive result: good generalizability -> stable effect
  - negative result: low generalizability -> sensitive effect
- **Exact replication**
  - positive result: effect does exist
  - negative result: ?
    - Was the replication really exact?
    - **Every deviation from the original study (except changes of test persons & investigators) could explain the negative result**

[5, 6]

# Problems of an Exact Replication (1)

---

## ■ Time

### ➤ Changes of measurement

- *Flynn-Effect*

### ➤ Changes in the effect of stimulus material

- *Researchers tried to induce fear in participants (1953) by telling them that improper care of their teeth might result in “cancer, paralysis or other secondary diseases”*

➤ **Today it might arouse disbelief rather than fear**

### ➤ Effect itself could change or diminish over time

- *People can change e.g. change of stereotypes, Flynn-effect*

[5, 6, 8]

## Problems of an Exact Replication (2)

---

- **Culture**
  - Cultural differences, subcultures or changes in culture
- **Power**
  - Is 80 percent sufficient?
- **Single replication study**
  - What does a single study over another single study tell us?
- **Regression to the mean**

[5, 6, 8]

# Regression to the Mean

---

- **Statistical phenomenon, that occurs whenever extreme values were measured.**
  - If they are measured again it is highly likely that the observed values are now closer to the mean of the population.
  - examples
    - *Blood pressure, sports*
    - *Placebo groups*
      - **Study results -> first study vs replication study**



Bland, J. M., & Altman, D. G. (1994). Regression towards the mean. *BMJ: British Medical Journal*, 308(6942), 1499.

Barnett, A. G., Van Der Pols, J. C., & Dobson, A. J. (2004). Regression to the mean: what it is and how to deal with it. *International journal of epidemiology*, 34(1), 215-220.

# Critique on the Replication Study (Open Sci. Collab., 2015)

---

## Comment on “Estimating the reproducibility of psychological science”

Citations of original study: 3672  
Citations of Comment: 369

Daniel T. Gilbert,<sup>1\*†</sup> Gary King,<sup>1</sup> Stephen Pettigrew,<sup>1</sup> Timothy D. Wilson<sup>2</sup>

*“We show that this article contains three statistical errors and provides no support for such a conclusion. Indeed, the data are consistent with the opposite conclusion, namely, that the reproducibility of psychological science is quite high.”*

[6]

# Critique on the Replication Study (Open Sci. Collab., 2015)

---

- **Many of OSC's replication studies drew their samples from different populations than the original studies did**
  - An original study that measured American's attitudes toward African-Americans was replicated with Italians, who do not share the same stereotypes
    - These infidelities are potential sources of random error that the OSC's benchmark did not consider
- **"In contrast, MLP [Many Labs Project] attempted to replicate each of its studies 35 or 36 times and then pooled the data. MLP's much more powerful method produced a much more heartening result:**
  - A full 85% of the original studies were successfully replicated."

[6]

# Investigating Variation in Replicability

## A “Many Labs” Replication Project

Richard A. Klein,<sup>1</sup> Kate A. Ratliff,<sup>1</sup> Michelangelo Vianello,<sup>2</sup> Reginald B. Adams Jr.,<sup>3</sup> Štěpán Bahník,<sup>4</sup> Michael J. Bernstein,<sup>5</sup> Konrad Bocian,<sup>6</sup> Mark J. Brandt,<sup>7</sup> Beach Brooks,<sup>1</sup> Claudia Chloe Brumbaugh,<sup>8</sup> Zeynep Cemalcilar,<sup>9</sup> Jesse Chandler,<sup>10,36</sup> Winnee Cheong,<sup>11</sup> William E. Davis,<sup>12</sup> Thierry Devos,<sup>13</sup> Matthew Eisner,<sup>10</sup> Natalia Frankowska,<sup>6</sup> David Furrow,<sup>15</sup> Elisa Maria Galliani,<sup>2</sup> Fred Hasselman,<sup>16,37</sup> Joshua A. Hicks,<sup>12</sup> James F. Hovermale,<sup>17</sup> S. Jane Hunt,<sup>18</sup> Jeffrey R. Huntsinger,<sup>19</sup> Hans IJzerman,<sup>7</sup> Melissa-Sue John,<sup>20</sup> Jennifer A. Joy-Gaba,<sup>17</sup> Heather Barry Kappes,<sup>21</sup> Lacy E. Krueger,<sup>18</sup> Jaime Kurtz,<sup>22</sup> Carmel A. Levitan,<sup>23</sup> Robyn K. Mallett,<sup>19</sup> Wendy L. Morris,<sup>24</sup> Anthony J. Nelson,<sup>3</sup> Jason A. Nier,<sup>25</sup> Grant Packard,<sup>26</sup> Ronaldo Pilati,<sup>27</sup> Abraham M. Rutchick,<sup>28</sup> Kathleen Schmidt,<sup>29</sup> Jeanine L. Skorinko,<sup>20</sup> Robert Smith,<sup>14</sup> Troy G. Steiner,<sup>3</sup> Justin Storbeck,<sup>8</sup> Lyn M. Van Swol,<sup>30</sup> Donna Thompson,<sup>15</sup> A. E. van ‘t Veer,<sup>7</sup> Leigh Ann Vaughn,<sup>31</sup> Marek Vranka,<sup>32</sup> Aaron L. Wichman,<sup>33</sup> Julie A. Woodzicka,<sup>34</sup> and Brian A. Nosek<sup>29,35</sup>

**Abstract.** Although replication is a central tenet of science, direct replications are rare in psychology. This research tested variation in the replicability of 13 classic and contemporary effects across 36 independent samples totaling 6,344 participants. In the aggregate, 10 effects replicated consistently. One effect – imagined contact reducing prejudice – showed weak support for replicability. And two effects – flag priming influencing conservatism and currency priming influencing system justification – did not replicate. We compared whether the conditions such as lab versus online or US versus international sample predicted effect magnitudes. By and large they did not. The results of this small sample of effects suggest that replicability is more dependent on the effect itself than on the sample and setting used to investigate the effect.

## Summary: Critique on the Replication Study

---

*“We applaud efforts to improve psychological science, many of which have been careful, responsible, and effective [...], and we appreciate the effort that went into producing OSC [Open Science Collaboration]. But metascience is not exempt from the rules of science. **OSC used a benchmark that did not take into account the multiple sources of error in their data, used a relatively low-powered design that demonstrably underestimates the true rate of replication, and permitted considerable infidelities that almost certainly biased their replication studies toward failure.***

***As a result, OSC seriously underestimated the reproducibility of psychological science.”***

[6]

# Why the crisis was needed

---

The Crisis of Confidence in Research Findings in Psychology: Is Lack of Replication the Real Problem? Or Is It Something Else?

Frank L. Schmidt  
University of Iowa

In-Sue Oh  
Temple University

*“Meta-analysis can provide the solution to these problems if the problems of publication bias and questionable research practices are successfully addressed. **The real problem is not a lack of replication; it is the distortion of our research literatures caused by publication bias and questionable research practices.**”*

# Summary of Problem Areas in Scientific Literature

---

- **False-Positive Errors**
  - Flexibility (degrees of freedom) in data analysis facilitates  $p$ -Hacking and HARKing
- **File-Drawer Problem**
- **Publication Bias**
- **Errors in use of selection, usage and interpretation of statistical methods**
  - optional stopping,  $p$ -values, ...
- **Power**
- **Replications**
- **Fraud**
- ...

# Summary of Problem Areas in Scientific Literature

---

- **False-Positive Errors**
  - Flexibility (degrees of freedom in data analysis facilitates  $p$ -Hacking and HARKing)
- **File-Drawer Problem**
- **Publication Bias**
- **Errors in use of selection, usage and interpretation of statistical methods**
  - optional stopping,  $p$ -values, ...
- **Power**
- **Replications**
- **Fraud**
- ...

*Why does  
this  
happen?*

# Why does this happen?

---

- **“Publish or Perish”**
  - the scientific system in general
- **Other interests than exploring the truth**
  - Interests of pharmaceutical companies, agribusiness, arms companies, ...
- **Human failures**
- ...



# Structure

---

- I. The Crisis
- II. Is there really a Crisis?
- III. Consequences & Changes





# III. Consequences & Changes

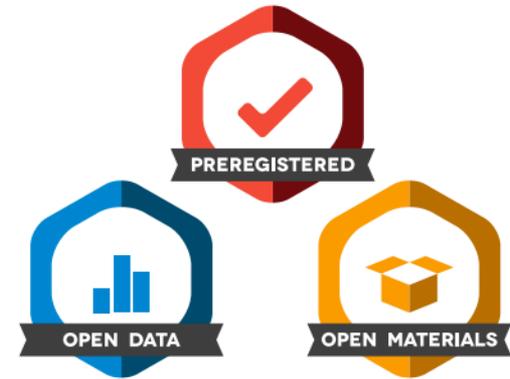
---

**(1) Solutions**

**(2) Future Directions**

# Solutions

---



- **Open Science**
  - Free access to the article (Open Access Journals)
    - But also to the data, the R-script, the materials, etc.
- **Meta-Analysis**
- **More replication projects**
- **Change of the culture**
  - less pressure to publish and more focus on quality
- **Change of statistical methods**
- **Pre-Registration**

# Change of Statistical Methods

---

- **Discussion of p-values**
  - Bayesian vs frequentist statistic
  - Never report p-values without effect sizes
- **A priori power analysis**
  - These changes would not decrease the motivation behind the p-hacking!
  - The culture of publish or perish is probably one of the deepest roots of the crisis.
  - **The pressure to publish has to be reduced in combination with an increase in the quality and the standards.**

# Pre-Registration & Registered Reports

## ■ Pre-Registration

- Registration of the hypothesis before analyzing the data
- No peer review & no guarantee to get published

## ■ Registered Reports



- After the 1. peer review high quality **protocols get provisionally accepted for publication** if the authors follow through with the registered methodology.

[9]

# Registered Reports

---



***"Because the study is accepted in advance, the incentives for authors change from producing the most beautiful story to the most accurate one."***

(Chris Chambers, Professor at Cardiff University)

[9]

# Registered Reports



## ■ Decreases

➤ the “need to get good results”

➤ reduces or even diminishes therefore the motivation behind p-hacking, HARKing, etc.

➤ Publication Bias

➤ File-Drawer-Problem

➤ Problems of low power

## ■ Increases

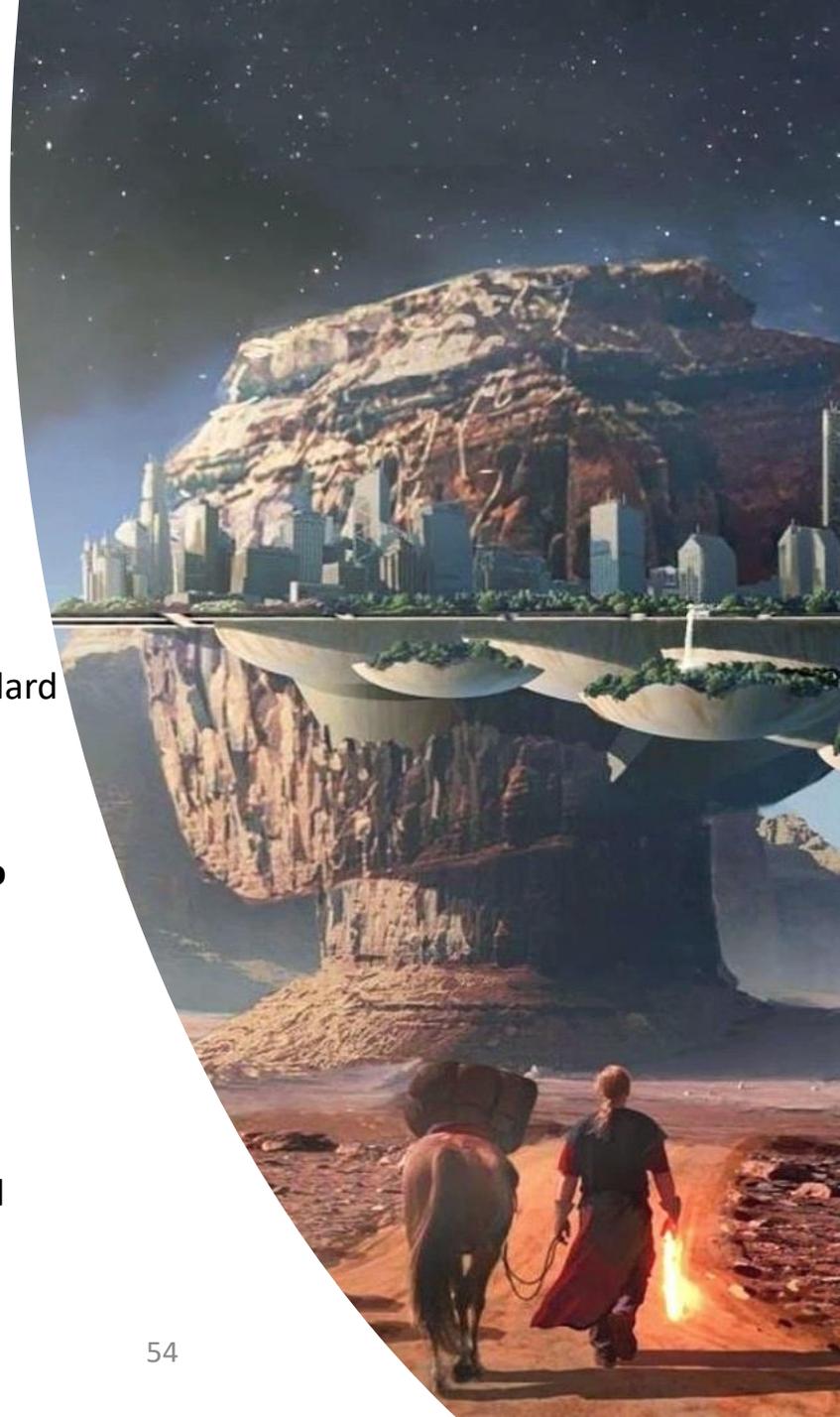
➤ quality (two steps of peer review)

■ **Saves money (critique takes place before the data collection not afterwards)**

[9]

# Future Directions

- **Crisis must be understood, discussed and resolved across disciplines**
- **New ideas & solutions must be transferred to the standard**
  - Registered Reports
  - The idea of open science should become the standard
    - **Scientific knowledge should be made available to all researchers worldwide**
- **Research should incorporate more feedback loops to help detect errors in the system more quickly**
- **Communication**
  - More interdisciplinary communication to identify similar problems and adopt solutions faster
  - The communication of research results should be better guided, that even persons without technical expertise have access and can understand the contents and their meaning correctly.





# “Never Waste a Good Crisis”

---

**Daniel J. Drucker**

## Sources

- [1] Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual review of psychology*, 69, 511-534.
- [2] Ioannidis John, P. A. (2017). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8).
- [3] 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.
- [4] Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- [5] Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, 9(1), 59-71.
- [6] Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the reproducibility of psychological science”. *Science*, 351(6277), 1037-1037.
- [7] Van Berkel, L., & Crandall, C. S. (2018). *Conceptual Replication of the Automaticity of Hierarchy: Multiple Methods to Test a Hypothesis*. SAGE Publications Ltd.
- [8] Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean?. *American Psychologist*, 70(6), 487.
- [9] <https://cos.io/rr/>

## Pictures:

[Background Image](#)

(Remaining images are also links for the image source)

