

**Before it will get better** ... (On the deepening replication crisis in the social sciences and what to do about it)



Andreas Ortmann, UNSW Australia Business School, Sydney

CARMA Workshop on Mathematical Aspects of behaviorural Economics and Finance University of Newcastle, November 14, 2015 As the old adage has it,

Before it will get better ...,

... it will get worse.



### Some terminology

p-hacking

=> Running a study until you get (statistical) significance (nevermind economic significance), an unfortunate consequence of the social sciences' obsession with NHST ... (recall talks by MA, SZ; see also work by Gigerenzer and various coauthors)

Not properly powering up a study

=> Rather than doing proper statistical power calculations beforehand (i.e., determining the appropriate risk of failure to reject a false null hypothesis/to detect a true effect, and then -- based on prior information about effect size -- computing the required sample size based on the choice of β and assumed effect size), using rules of thumbs for sample size.

Publication bias

=> Studies that report no effect not being published (e.g., Simonsohn et al. JEP:G 2014; Ferguson & Heene PoPS 2012; Farelli SM 2012; earlier Rosenthal PB 1979, and many others; see also Head et al. PLoS 2015)



### Replication crisis is not a privilege of the social sciences

The extent and consequences of p-hacking, and publication biases in science, are well documented for the sciences more generally:

#### Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 + DOI: 10.1371/journal.pmed.0020124

#### Abetract

#### Summary

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 among the relationships probed in each scientific field. In this framework, 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. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Citation: Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124



# Replication crisis is not a privilege of the social sciences, ctd.

The extent and consequences of p-hacking (major theme in Simonsohn's work), and publication biases in science, are well documented for the sciences more generally:

### The Extent and Consequences of P-Hacking in Science

Megan L. Head , Luke Holman, Rob Lanfear, Andrew T. Kahn, Michael D. Jennions

Published: March 13, 2015 • DOI: 10.1371/journal.pbio.1002108

Abstract

A focus on novel, confirmatory, and statistically significant results leads to substantial bias in the scientific literature. One type of bias, known as "p-hacking," occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant. Here, we use text-mining to demonstrate that p-hacking is widespread throughout science. We then illustrate how one can test for p-hacking when performing a meta-analysis and show that, while p-hacking is probably common, its effect seems to be weak relative to the real effect sizes being measured. This result suggests that p-hacking probably does not drastically alter scientific consensuses drawn from meta-analyses.

Citation: Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The Extent and Consequences of P-Hacking in Science. PLoS Biol 13(3): e1002106. doi:10.1371/journal.pbio.1002106



### Investigating Variation in Replicability: A "Many Labs" Replication Project

Public 👁 14 🖸 2

Contributors: Richard A. Klein (/tmwry/), Kate Ratliff (/daang/), Michelangelo Vianello (/d42gk/), Reginald B. Adams, Jr., Štěpán Bahník (/bga2x/), Michael Jason Bernstein (/7nbzd/), Konrad Bocian (/p97qc/), Mark Brandt (/w7frf/), Beach Brooks (/2x4da/),

Claudia Brumbaugh (/jv8bn/), Zeynep Cemalcilar (/e6wxt/), Jesse J. Chandler (/ubit5/), Winnee Cheong (/2y6wh/), William E. Davis (/5zjt7/),

Date created: 2013-06-15 05:29 AM | Last Updated: 2015-09-25 07:55 AM

Category: Project 📦

Description: We conducted replications of 13 effects in psychological science with 36 samples and more than 6000 participants. We examined heterogeneity in replicability across sample and setting.



Judgment and Decision Making, Vol. 10, No. 1, January 2015, pp. 1-17

On making the right choice: A meta-analysis and large-scale replication attempt of the unconscious thought advantage

Mark R. Nieuwenstein<sup>\*</sup> Tjardie Wierenga<sup>#</sup> Richard D. Morey<sup>#</sup> Jelte M. Wicherts<sup>§</sup> Tesse N. Blom<sup>§</sup> Eric-Jan Wagenmakers<sup>§</sup> Hedderik van Rijn<sup>§</sup>

Are difficult decisions best made after a momentary diversion of thought? Previous research addressing this important question has yielded dozens of experiments in which participants were asked to choose the best of several options (e.g., cars or apartments) either after conscious deliberation, or after a momentary diversion of thought induced by an unrelated task. The results of these studies were mixed. Some found that participants who had first performed the unrelated task were more likely to choose the best option, whereas others found no evidence for this so-called unconscious thought advantage (UTA). The current study examined two accounts of this inconsistency in previous findings. According to the reliability account, the UTA does not exist and previous reports of this effect concern nothing but spurious effects obtained with an unreliable paradigm. In contrast, the moderator account proposes that the UTA is a real effect that occurs only when certain conditions are met in the choice task. To test these accounts, we conducted a meta-analysis and a large-scale replication study (N = 399) that met the conditions deemed optimal for replicating the UTA. Consistent with the reliability account, the large-scale replication study yielded no evidence for the UTA, and the meta-analysis showed that previous reports of the UTA were confined to underpowered studies that used relatively small sample sizes. Furthermore, the results of the large-scale study also dispelled the recent suggestion that the UTA might be gender-specific. Accordingly, we conclude that there exists no reliable support for the claim that a momentary diversion of thought leads to better decision making than a period of deliberation.



Behav Brain Sci., 2014 Feb;37(1):1-19. doi: 10.1017/S0140525X12003214. Epub 2014 Jan 24.

#### Unconscious influences on decision making: a critical review.

Newell BR<sup>1</sup>, Shanks DR<sup>2</sup>.

Author information

#### Abstract

To what extent do we know our own minds when making decisions? Variants of this question have preoccupied researchers in a wide range of domains, from mainstream experimental psychology (cognition, perception, social behavior) to cognitive neuroscience and behavioral economics. A pervasive view places a heavy explanatory burden on an intelligent cognitive unconscious, with many theories assigning causally effective roles to unconscious influences. This article presents a novel framework for evaluating these claims and reviews evidence from three major bodies of research in which unconscious factors have been studied: multiple-cue judgment, deliberation without attention, and decisions under uncertainty. Studies of priming (subliminal and primes-to-behavior) and the role of awareness in movement and perception (e.g., timing of willed actions, blindsight) are also given brief consideration. The review highlights that inadequate procedures for assessing awareness, failures to consider artifactual explanations of "landmark" results, and a tendency to uncritically accept conclusions that fit with our intuitions have all contributed to unconscious influences being ascribed inflated and erroneous explanatory power in theories of decision making. The review concludes by recommending that future research should focus on tasks in which participants' attention is diverted away from the experimenter's hypothesis, rather than the highly reflective tasks that are currently often employed.



full text

The Open Science Collaboration:

- almost 300 researchers that tried to replicate
- the results of 100 papers published in three leading psychology journals (2008)
   Trying to do direct replications, the researchers did not succeed in the clear majority of cases and on average they found the mean effect size to be only half of what was reported in the original studies.



## 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.



The Open Science Collaboration:

- Trying to do direct replications, the researchers did not succeed in the clear majority of cases and on average they found the mean effect size to be only half of what was reported in the original studies.
  - There are important questions about this report (see also http://daniellakens.blogspot.nl/2015/08/power-of-replications-in.html):

Se Character & Context

#### Simone Schnall on her Experience with a Registered Replication Project

Newsletter Twitter Primer

Facebook Who We Are



UNSW THE UNIVERSITY OF NEW SOUTH WALES

🛔 Simone Schnall 🛛 🖉 May 23, 2014 📲 In the Journals 🌘 71 Comments

This post originally appeared <u>here</u> and is re-posted in its entirety below. Brent Donnellan, one of the authors of the attempted replication, has written a related post <u>here</u>. For more insights into the new special replication issue of Social Psychology, see our weekly link round-up <u>here</u>.

That said, there is a growing insight among psychologists that a corner stone of the scientific edifice is in serious need of restauration (replicability etc.) and thinking about the way one does experiments is as a good a start as any:

Zeistgeist: Psychological Experimentation, Cognition, Language, and Academia.

Friday, August 16, 2013

#### 50 Questions About Messy Rooms and Clean Data

About a month ago, I had a difficult conversation with my daughter. Her year in college had not gone particularly well and I asked her what she was going to do differently next year. One of the first things she was going to do, she said, was to clean up her student room. It was just too cluttered to concentrate.



About Me

Rolf Zwaan G+ Follow 189

View my complete profile

### Here's Why That Study Claiming Religious Kids Are Less Altruistic Stinks. Updates

NOVEMBER 9, 2015 / BRIGGS / 151 COMMENTS



Most recently, two Federal Reserve economists Chang, Andrew C., and Phillip Li (2015)

"Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not", Finance and Economics Discussion Series 2015-083. Washington: Board of Governors of the Federal Reserve System, http://dx.doi.org/10.17016/FEDS.2015.083.

came to the alarming conclusion that economics research is "usually not" replicable.. Attempt at replicating 67 empirical papers in 13 reputable academic journals

- without assistance by the original researchers could replicate only a third of it
- with assistance ... percentage increased to about half of the original results

The failure to replicate the majority of studies, which is based on an **attempt at replicating past results with existing data sets rather than with newly produced data**, is interesting by itself, as it indicates that the reporting practices and requirements are seriously deficient. (This brings us back to some of the themes that DB and JB brought up already yesterday.)



Ignoring outright fraudulent behavior (Stapel, Sanna, Smeesters, Foerster, Zaman, Lichtenthaler, ...), good science is messy and actually hard work and reasonable people can reasonably disagree.

Case in point: just-published study by Silberzahn & Uhlmann (S&U), two researchers who got engaged in methodological debates when Uri Simohnson [http://opim.wharton.upenn.edu/~uws/] questioned the results of an earlier S&U study that suggested that noble-sounding German names could boost careers.

- Re-running the analysis with a better analytical approach, Simonsohn did not confirm some such effect.
- S&U ended up conceding that much in a joint paper with Simonsohn.
- Simonsohn and S&U ought to be applauded for their collaboration and in the case of S&U their public concession that their methods were wanting.
- There is nothing to be ashamed about (and indeed it can lead to interesting insights).



Case in point: just-published study by Silberzahn & Uhlmann,

[http://www.nature.com/news/crowdsourced-research-many-hands-make-tight-work-<u>1.18508</u>] In their new study. S&U provided a data set and then asked more than two dozen teams of researchers to determine whether skin color of a soccer players across four major leagues (yes, ....) influenced the frequency with which a red card would be issued.

- Somewhat shockingly, the answers were rather diverse.
- Of the 29 teams, 20 found a statistically significant correlation with the median suggesting that dark-skinned players were 1.3 times more likely than light-skinned players to be sent off.
- But, "findings varied enormously, from a slight (and non-significant) tendency for referees to give more red cards to light-skinned players to a strong trend of giving more red cards to dark-skinned players."
- Interestingly, this diversity of results survived even an iteration of methodological discussion among all researchers. Did I mention that good science is messy and actually hard work and reasonable people can reasonably disagree?



Case in point: just-published study by Silberzahn & Uhlmann:

Even under the best of circumstances - one data set, what seems like a straightforward question to answer, and exchange of ideas among a set of researchers about the best method - to arrive at consensus can be extraordinarily difficult.

Even more difficult if the presence, or absence of, effects is identified through multiple data sets compiled empirically (experimentally) by competing teams of researchers who have incentives to add to the "Vast Graveyard of Undead Theories" (Ferguson & Heene, PoPS 2012).

Tantalizing evidence that many EE studies are severely under-powered:

#### Exploring the Meaning of Significance in Experimental Economics

Le Zhang Curtin University - School of Economics and Finance

Andreas Ortmann UNSW Australia Business School, School of Economics

November 16, 2013

UNSW Australian School of Business Research Paper No. 2013-32

#### Abstract:

Null Hypothesis Significance Testing has been widely used in the experimental economics literature. Typically, attention is restricted to type-I-errors. We demonstrate that not taking type-II errors into account is problematic. We also provide evidence, for one prominent area in experimental economics (dictator game experiments), that most studies are severely underpowered, suggesting that their findings are questionable. We then illustrate with several examples how poor (no) power planning can lead to questionable results.



Also (and on the positive side):

### **Behavioral Economics Replication Project**

Prediction markets have recently been established as popular forecasting tools. By organizing markets on research projects, we are exploring how such markets can contribute to the objectives of scientific research.

Learn more » (/repoverview.html)

#### **Project overview**

This project will provide evidence of how accurately peer prediction markets can forecast replication of scientific experiments in economics.

In order to incentivize prediction market activity, and collect evidence on actual replication, eighteen (18) prominently published studies in experimental economics were chosen for trading in prediction markets, followed by replication. They are laboratory studies, using student participants, that were published in the American Economic Review (AER) or in the Quarterly Journal of Economics (QJE) in the years 2011 to 2014, testing specific hypotheses using between-subjects designs.



<sup>©</sup> The Science Prediction Market Project (mailto:predictionreplication@hhs.se), 2015

- 1. Start every project with a proper literature review. Referees and editors need to become more adamant about proper acknowledgment of what is out there. This is a real issue and it's becoming a pet peeve of mine (e.g., recent *JEP* issue on overconfidence)
- 2. We need clearer reporting and data collection conventions (a 21 word solution?):

#### Richard Gill

Original talk December 2012; slides updated March 2013; postscript July 2013

#### http://www.math.leidenuniv.nl/~gill

- Scientific = Reproducible: Data preparation and data analysis are integral parts of experiment
- Keeping proper log-books of all steps of data preparation, manipulation, selection/exclusion of cases,
   makes the experiment reproducible
- Sharing statistical analyses over several authors is almost necessary in order to prevent errors

- Data collection protocol should be written down in advance in detail and followed carefully
- Exploratory analyses, pilot studies ... also science
- Replicating others' experiments: also science
- It's easy to make mistakes doing statistical analyses: the statistician needs a co-pilot
- Senior co-authors co-responsible for good scientific practices of young scientists in their group



3. We need a reduction of the wiggle room that is currently afforded:

False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant.

Simmons JP<sup>1</sup>, Nelson LD, Simonsohn U.

Show more

A study (and if necessary its online appendices ought to have enough information to allow replication without others - specifically the original authors -- having to be consulted (see discussion of Chen & Li study, Silberzahn & Uhlmann study, various of the psych studies)

Open question to my mind if it is good/best practice to invite original authors ...

Data - lest they are confidential - ought to be deposited with the journal in which the articles was published; dito for various programming files etc.

The replication recipes currently being tested are good way to get started:

The Replication Recipe: What makes for a convincing replication?

```
Mark J. Brandt<sup>e, 1</sup>, Hans IJzerman<sup>e, 1</sup>, Ap Dijksterhuis<sup>b, 2</sup>, Frank J. Farach<sup>e, 2</sup>, Jason Geller<sup>4, 2</sup>,
Roger Giner-Sorolla<sup>e, 2</sup>, James A. Grange<sup>1, 2</sup>, Marco Perugini<sup>g, 2</sup>, Jeffrey R. Spies<sup>b, 2</sup>, Anna van 't Veer<sup>e, 1, 2</sup>
Under a Creative Commons license
```



4. We need more meta-studies to inform proper design and implementation of individual studies (especially in economics). Seriously. Management apparently, not so much ...

#### An Assessment of the Magnitude of Effect Sizes: Evidence From 30 Years of Meta-Analysis in Management

Journal of Leadership & Organizational Studies I–16 © The Authors 2015 Reprints and permissions: sagepub.com/journalsPermissions.nav DOI: 10.1177/1548051815614321 jlo.sagepub.com

#### Ted A. Paterson<sup>1</sup>, P.D. Harms<sup>2</sup>, Piers Steel<sup>3</sup>, and Marcus Credé<sup>4</sup>

#### Abstract

This study compiles information from more than 250 meta-analyses conducted over the past 30 years to assess the magnitude of reported effect sizes in the organizational behavior (OB)/human resources (HR) literatures. Our analysis revealed an average uncorrected effect of r = .227 and an average corrected effect of  $\rho = .278$  ( $SD\rho = .140$ ). Based on the distribution of effect sizes we report, Cohen's effect size benchmarks are not appropriate for use in OB/HR research as they overestimate the actual breakpoints between small, medium, and large effects. We also assessed the average statistical power reported in meta-analytic conclusions and found substantial evidence that the majority of primary studies in the management literature are statistically underpowered. Finally, we investigated the impact of the file drawer problem in meta-analyses and our findings indicate that the file drawer problem is not a significant concern for meta-analysts. We conclude by discussing various implications of this study for OB/HR researchers.



5. We need more adversarial collaborations of the kind Mellers et al. (PS 2002) have or Blavatskyy et al. (2015, under review) have done:

### Now you see it, now you don't: *How to make* the Allais paradox appear, disappear, or reverse

Pavlo Blavatskyy<sup>1</sup>, Andreas Ortmann<sup>2</sup>, and Valentyn Panchenko<sup>3</sup> \*

Abstract: The Allais paradox, or the common-consequence effect, is arguably the best-known behavioral regularity in individual decision making under risk. A common perception in the literature, which motivated the development of numerous generalized non-expected utility theories, is that the Allais paradox is a robust empirical finding. We argue that such a perception does not accurately reflect the existing experimental evidence on the Allais paradox and show how specific choices of design and implementation characteristics and parameters can make the effect appear, disappear, or reverse. For example, our results suggest that the Allais paradox is likely to disappear when lotteries involve relatively small outcomes under real financial incentives and probability distributions are described as compound lotteries or in a frequency format (rather than as reduced-form simple lotteries). We also find that the Allais paradox is likely to get reversed when lotteries are designed with an even division of the probability mass between the lowest and the highest outcomes.



- 6. We need more studies of the kind that Silberzahn & Uhlmann have done.
- 7. We need more systematic (tournament) studies of the kind that Erev et al have done.

#### From Anomalies to Forecasts: Choice Prediction Competition for Decisions under Risk and Ambiguity (CPC2015)

Supported by the Max Wertheimer Minerva Center for Cognitive Processing and Human Performance Organized by: Ido Erev, Eyal Ert, and Ori Plonsky Submission deadline: May 17th, 2015 | Early registration until April 1st, 2015

News and updates

Motivation and basic idea

Registration & Submission

Problem Space

Problem Selection Algorithm

Method

Aggregated Results

Experiment 1: Replication of well-known

decision phenomena Experiment 2:

Estimation Set

Experiment 3:

Competition set

Raw Data

Competition Rules

Baseline Models

Competition results

#### Motivation and the basic idea

Experimental studies of human choice behavior have documented clear violations of rational economic theory and triggered the development of behavioral economics. Yet, the impact of these careful studies on applied economic analyses, and policy decisions, is not large. One justification for the tendency to ignore the experimental evidence involves the assertion that the behavioral literature highlights contradicting deviations from maximization, and it is not easy to predict which deviation is likely to be more important in specific situations.

To address this problem Kahneman and Tversky (1979) proposed a model (Prospect theory) that captures the joint effect of four of the most important deviations from maximization: the certainty effect (Allais paradox, Allais, 1953), the reflection effect, overweighting of low probability extreme events, and loss aversion (see top four rows in Table 1). The current paper extends this and similar efforts (see e.g., Thaler & Johnson, 1990; Brandstätter, Gigerenzer, & Hertwig, 2006; Birnbaum, 2008; Wakker, 2010; Erev et al., 2010) by facilitating the derivation and comparison of models that capture the joint impact of the four "prospect theory effects" and ten additional phenomena (see Table 1).

These choice phenomena were replicated under one "standard" setting (Hertwig & Ortmann, 2001): choice with real stakes in a space of experimental tasks wide enough to replicate all the phenomena illustrated in Table 1. The results suggest that all 14 phenomena emerge in our setting. Yet, their magnitude tends to be smaller than their magnitude in the original demonstrations.

The current choice prediction competition focuses on developing models that can capture all of these phenomena but also predict behavior in other choice problems. To calibrate the models we ran an "estimation set" study that included 60, randomly selected, choice problems.

A draft that summarizes the results of the estimation set can be found here: CPC2015 paper (January 15, 2015)

- 8. We need to name and shame those journals that publish sensationalist studies but then refuse to publish failure to replicate (e.g., Gneezy et al. QJE 2006; Rydval et al. EE 2009 but many many others - how do you effect some such change?).
- 9. We need more initiatives (e.g., pre-registered studies) such as those by

Journal of the Economic Science Association, Experimental Economics, Journal of Experimental Social Psychology, Journal of Personality and Social Psychology, Psychological Science, Perspectives on Psychological Science....

- 10. We need more appreciation of proper econometrics
- 11. We need more thinking about generalizability
- 12. We need an understanding that a failed replication in many instances constitutes real progress and is nothing to be ashamed about



#### Summarizing ...

As the old adage has it,

Before it will get better ...,

... ít will get worse.

I hope that the current methodological debates will strengthen the social sciences in the long run, as they will up the ante on what it takes for a study to provide informational value (to be valid) and to be reliable.

- Because the considerable damage it can inflict on individuals' productivity and reputation [http://www.spspblog.org/simone-schnall-on-her-experience-with-a-registered-replication-project/], it will also, hopefully, provide incentive to clarify what minimal reporting standards and acceptable replication etiquettes are.
- I hope that the current methodological debates will give incentives for journals to become serious about their data set collection efforts although it has to be noted that some data sets have been provided under confidentiality provisions that prohibit such posting.



Summarizing, ctd ...

As the old adage has it,

Before it will get better ...,

... ít will get worse.

It would be good if the current debate would be conducted calmly and cooly. To call the proponents of replication studies a bunch of self-righteous, self-appointed sherrifs, or replication police, or worse, is as unproductive as is abandoning the presumption of innocence.

Accusations of misbehaviour ought to be carefully avoided and preferably sorted out through the appropriate non-public channels, or privately (as in the case of S&U and Simonsohn), rather than social media which - while entertaining - seem singularly poorly suited to decide on scientific merit.



#### Just in case you are in, or nearby, Sydney on November 27

#### 2015 workshop in Experimental Methods: The replicability crisis in the social sciences and how to address it

The BizLab at UNSW is hosting a half-day Workshop in Experimental Methods for Research in Social Science and Business on Friday, November 27th.

It is no longer news that both economics and psychology face serious replication crises that seem to be brought about – to a considerable extent – by questionable practices such as p-hacking. Since replication is an important component of cumulative science, these crises pose a major challenge and are of considerable interest to everyone working experimentally or empirically in accounting, economics, finance, management, marketing, psychology and other related areas.

The workshop is meant to offer an up-to-date assessment of the current state of the crises and strategies to overcome them.

We welcome academics from across the business disciplines and social sciences who are interested in experimental methods and interdisciplinary exchange. It is anticipated both faculty and PhD students from across Australia will attend the workshop, and there will be plenty of time for informal networking.

#### Details

. . .

- Date: Friday, November 27th
- Time: 12pm 6.30pm, followed by drinks and fingerfood
- Location: UNSW Kensington Campus, Sydney
   Ainsworth Building (J17) Theatre G03
- Catering: Lunch, coffee, refreshments, drinks, and fingerfood will be provided free of charge

#### Registration

Registration is required. No registration, no participation. Registration deadline is November 24, 2015 17:00 AEDT





# Thank you!

