

## ¶ Training: Transparent and Reproducible Research

---

Aleksandr Michuda

Center for Data Science for Enterprise and Society, Cornell University

Special Thanks to Oscar Barriga-Cabanillas and Matthieu Stigler!

- Welcome!
  - **Transparency and Reproducibility in Research (You are here.)**
  - Data Management
  - Dynamic Documents

# Transparency and Reproducibility in Research

- Is there a problem with replicability and statistical validity?
- Is there a reproducibility crisis?
- How can we counter-act this?
- Replicability
  - Not a problem of results being intentionally falsified, although there are cases
  - There is a lack of incentives to replicate / validate
- Statistical Validity
  - P-hacking
  - Multiple testing
  - Workflow
  - Null results are not a failure

May 29, 2015 12:34 p.m.

## The Case of the Amazing Gay-Marriage Data: How a Graduate Student Reluctantly Uncovered a Huge Scientific Fraud

By Jesse Singal

NEW YORK



# Replication Disincentives

- The persistence of David Broockman led to the discovery of data falsification
- Michael LeCour falsified the data of his thesis:
  - Subjects interviewed about their acceptance of gay marriage
  - Treatment: Revealing interviewer's sexual orientation before asking questions
- LeCour received an offer from Princeton because of his job as thesis, which was published in *Science* in 2014
- Received media coverage; organizations changed their policies because of this
- Many people were suspicious, but the inclusion of names recognized in the publication acted as shielding - people wanted to believe them

# Replication Disincentives

- How did Broockman find out?
- Wanted to write a similar study; replication
- He first wanted to replicate it
- The cost and logistics were not within the reach of a grad student
  - LeCours would have to have a budget of over \$1,000,000 to carry the design
    - 10,000 respondents paid \$100 a piece
- Things unravelled from there...
- His own study didn't end up being consistent
- Raw data was too “orderly”

# So What?

- No incentive to carry out replications
- But there isn't much reward to questioning results
  - Either “shamelessly taking down a big name.”
  - Or being unfair/critical to peers
- Until recently, not really considered “science”
- Points to structural issue
- How do we change it?

Journal of Comments and Replications in Economics

PEER-REVIEWED. OPEN ACCESS. NO AUTHOR FEES.

## But Also...

- That's an extreme example
- Biggest problems are the things we do as part of standard research practices
  - Cutting/subsetting/"exploring" sample until we find an "interesting" (read: stat. sig.) result
    - Or rather, not being clear about when cutting is part of our *a priori* understanding of the model vs. exploratory work
  - Not considering null results as interesting
  - Not commenting code/making it easier for someone to read/run our code
  - Not taking version control seriously  
(paper\_1\_new\_final\_final2\_changed\_final12.docx)



## Preview of Solutions

Cutting/subsetting/“exploring” sample until we find an “interesting” (read: stat. sig.) result

**i** Solution:

Multiple Hypothesis Testing; Pre-analysis Plans

Not considering null results as interesting

### **i** Solution

Registered Reports (Also, the idea of *precise zeros*)

Not commenting code/making it easier for someone to read/run our code

### Solution

dynamic documents, code commenting, docker

# Preview of Solutions

Not taking version control seriously

**i** Solution

git

The problem:

- The publication of null effects is disappearing over time, in all disciplines. (Fanelli 2011) .
- Valuable information is being lost

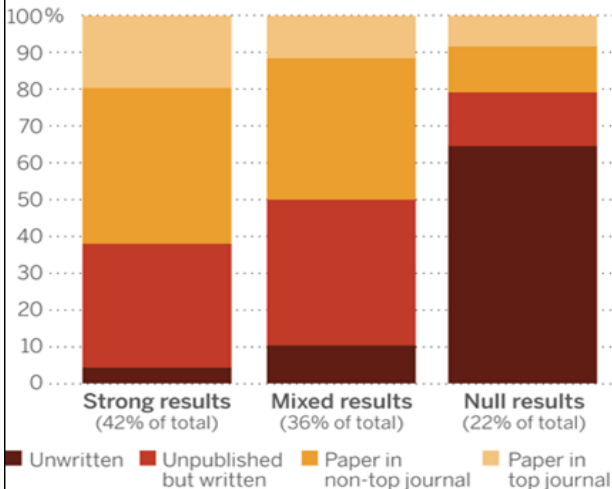
Null results are archived

- They are not published
- They are not registered

Nulls are needed to understand full distribution of results!

## Most null results are never written up

The fate of 221 social science experiments



Source: A. Franco *et al.*, *Science* (28 August)

# Registered Reports

- AKA Registered Reports, change time to peer review before the data collection, analysis and results.
  - Design a study
  - Send to a journal
  - Review based on the importance of the question and design quality
  - Get acceptance
  - Run the study, and publish even with null results

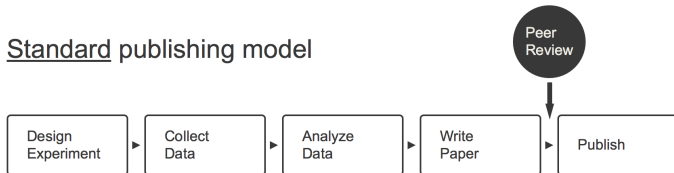
Hundreds of Journals Participating



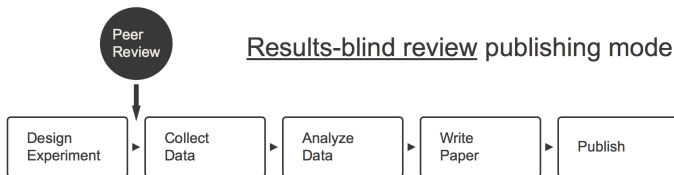


# What's the Difference?

## Standard publishing model



## Results-blind review publishing model



- And if that doesn't work, pre-register your studies!
- Gives credibility to work that you're doing
- Puts you “on the map” as having thought of the idea
- Can even do it for secondary data

- A null result can be interesting too!
- How to make sure that a null result is actually **0**?
  - Review design
  - Check that no-reject is not due to a small sample with respect to the effect size

**i** Definition (Minimum detectable effect size)

The smallest actual effect, which can be detected for a certain level of power and statistical significance.

- Consequences of always looking for p-values  $< 0.05$ :
  - Fraud
  - p-hacking: systematic search for significant results
  - unconscious bias: accept without criticism  $H_1$ , examine  $H_0$  meticulously
- This usually happens through constant testing of hypotheses

## Multiple Testing

- False Positive: Rejecting  $H_0$  when it is in fact not significant
- What is the probability of a false positive? Type I error:  $\alpha$  (usually 5%)
- What about with  $m$  hypotheses?
  - $R_i$  rejecting null hypothesis  $i$

$$P\{\text{at least one } H_i^0 \text{ rejected} \mid \text{all } H_i^0 \text{ correct}\}$$

$$= 1 - P\{\text{No rejections}\}$$

$$= 1 - P\{\bar{R}_1 \cap \bar{R}_2 \cap \dots \cap \bar{R}_m \mid H_i^0\}$$

$$= 1 - (1 - \alpha)^m \quad \text{if independent}$$

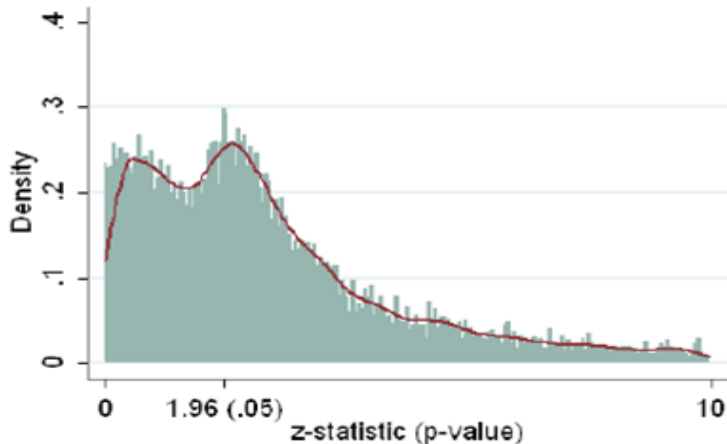
M=# tests:	1	2	3	5	10	20	50
	5%	10%	14%	22%	40%	64%	92%

## **i** P-Hacking

Systematic search for significant results on multiple combinations of i) dependent variables ii) combinations of independent variables; reporting only significant results. If we test enough hypotheses it is certain that we can reject at least one hypothesis, even if it is a false positive!!

## P-hacking is Real

(b) Unrounded distribution of z-statistics.





# P-hacking is Real

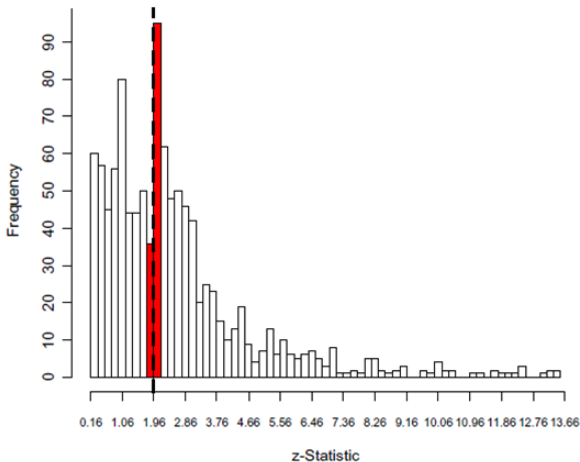


Figure 1(a). Histogram of  $z$ -statistics, *APSR* & *AJP* (Two-Tailed). Width of bars (0.20) approximately represents 10% caliper. Dotted line represents critical  $z$ -statistic (1.96) associated with  $p = 0.05$  significance level for one-tailed tests.

## Multiple Hypothesis Testing

$$\begin{aligned}P\{\#R_i \geq 1 | H_i^0 \forall i\} &= 1 - P\{\text{No rejections}\} \\&= 1 - P\{\bar{R}_1 \cap \bar{R}_2 \cap \dots \cap \bar{R}_m | H_i^0\} \\&= 1 - (1 - \alpha)^m \quad \text{if independent}\end{aligned}$$

The Bonferroni inequality:

$$P\{\#R_i \geq 1 | H_i^0 \forall i\} \leq m \cdot \alpha$$

# Multiple Hypothesis Testing

	Number of Tests					
	1	2	3	5	10	20
False Positive	5%	10%	14%	22%	40%	64%
No Adjustment	5%	5%	5%	5%	5%	5%
Bonferroni	5%	2.5%	1.67%	1.25%	1%	0.8%

# Multiple Hypothesis Testing

- Controlling for false positives:
  - Bonferroni: use the Boolean inequality  $P \{ \} < \alpha \cdot m \Rightarrow$  use  $\alpha = \frac{\alpha}{m}$ .

Problem: it is very conservative!

With  $m = 50$ ,  $\alpha = 0.001$  is used instead of 0.05!

Holmes: iterative procedure, not as conservative than Bonferroni

Control of the false discovery rate, the proportion of false positives:  
Benjamini – Hochberg

# When to use it?

- When to Use multiple corrections?
  - Tables with multiple results
  - After having explored many combinations of models
- Connects to idea of pre-analysis plans
  - plan for the number of tests you'll do so that the correction doesn't cause too much of a penalty

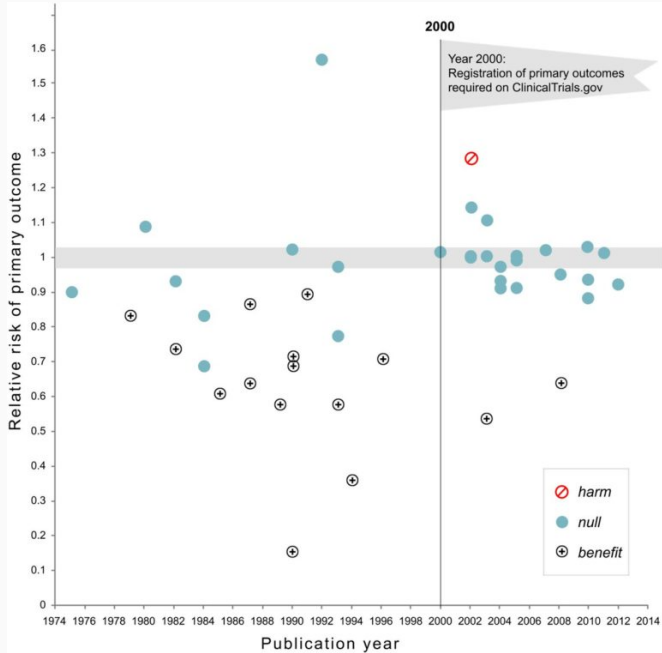
## **i** Pre-analysis Plan

Detailed description of the analyses to be conducted (hypotheses, variables, equations, controls, etc): This description is written before viewing the data.

By specifying the details before viewing the results, the plan is a safeguard against p-hacking.

## Where do you register?

- AEA Registry, actualmente solo para RCTs.
  - <https://socialscienceregistry.org>
- EGAP
  - <https://egap.org/design-registration>
- 3ie
  - <https://ridie.3ieimpact.org>
- Open Science Framework
  - <https://osf.io>





In a pre-analysis plan, a very clear distinction is made between two types of analysis:

- Confirmatory analysis : test the pre-registered hypotheses. If it is well done, it balances the false positive problem and which hypotheses are most important.
- Exploratory analysis : results not pre-registered. They come in a different section, and have less credibility (could be false positives).

# Research Transparency and Reproducibility Training (RT2) -- Virtual, 2020

[WUB](#) [PUB](#) [F](#) [O](#)

Contributors: Aleksandar Bogdanoski, Katie Hoebeling, Fernando Hoces de la Guardia, Edward Miguel, Tim Dennis, Benjamin Daniels, Katherine Kozlar, Graeme Blair, Katherine E. Kozlar, Cecilia Hyunjung Mo, **Aleksandr Michuda**, Denia Plammons, Jennifer Sturdy, Luiza Cardoso de Andrade, Danielle Kane, Vanessa Navarro Rodriguez, Luis Eduardo San Martin, Lars Vilhuber, Daniel Jacob Benjamin, Reid Otsuji

Date created: 2020-08-26 06:03 AM | Last Updated: 2020-10-01 05:56 PM

Identifier: DOI 10.17605/OSF.IO/A9HCK

Category: Project

Description:

Project page for the BITSS Research Transparency & Reproducibility Training (RT2), hosted online on Sep. 21-25, 2020.

License: CC0 1.0 Universal

Wiki

Important links:

- [Agenda](#)
- [Event page](#)
- [Participant Manual](#)

Files

Click on a storage provider or drag and drop to upload

Name	Modified
Research Transparency and Reproducibility Training (RT2) -- Virtual...	
<ul style="list-style-type: none"> <li> OSF Storage (United States)</li> <li> Day 1           <ul style="list-style-type: none"> <li> OSF Storage (United States)               <ul style="list-style-type: none"> <li> Aleksandar Bogdanoski -- BITSS Scholcom projects JDE reg... 2020-09-21 10:09 AM</li> <li> Introduction_Housekeeping_KatieHoebeling.pdf 2020-09-18 09:37 PM</li> <li> RT2-Virtual_Pre-registration-PAPs_Miguel_2020-09-21.pdf 2020-09-20 05:23 PM</li> <li> RT2-Virtual_Scientific-ethos_Miguel_2020-09-21.pdf 2020-09-20 05:23 PM</li> </ul> </li> </ul> </li> <li> Day 2           <ul style="list-style-type: none"> <li> OSF Storage (United States)               <ul style="list-style-type: none"> <li> Cecilia Mo -- PAPs for Observational Research.pdf 2020-09-22 11:17 AM</li> <li> PAP example from observational PAPs session.pdf 2020-09-22 01:36 PM</li> <li> PAP Starter -- Graeme Blair.Rmd 2020-09-21 11:33 PM</li> <li> Selecting a Research Design Slides - Graeme Blair.pdf 2020-09-22 10:42 AM</li> </ul> </li> </ul> </li> </ul>	

Citation

Components

- Day 1  
Bogdanoski, Hoebeling, Hoces de la Guardia & 17 more
- Day 2  
Bogdanoski, Hoebeling, Hoces de la Guardia & 17 more
- Day 3  
Bogdanoski, Hoebeling, Hoces de la Guardia & 17 more
- Day 4  
Bogdanoski, Hoebeling, Hoces de la Guardia & 17 more
- Day 5  
Bogdanoski, Hoebeling, Hoces de la Guardia & 17 more
- Office Hours & Other Materials  
Bogdanoski, Hoebeling, Hoces de la Guardia & 17 more
- Referenced materials  
Bogdanoski, Hoebeling, Hoces de la Guardia & 17 more

Tags

Add a tag to enhance discoverability

- “One-stop Shop” for Research
- Provides Integration with Github, cloud services
- Keeps track of your project and files changing
- **An easy place to write Pre-Analysis Plans and Registered Reports!**

- Take the time now to start a registered report, pre-registration or pre-analysis plan for your own project.