The Replication Crisis, Reproducibility, and the Reproducibility Project in Psychology

J. Edward Swan II

Mississippi State University

Monday, 16 October 2023

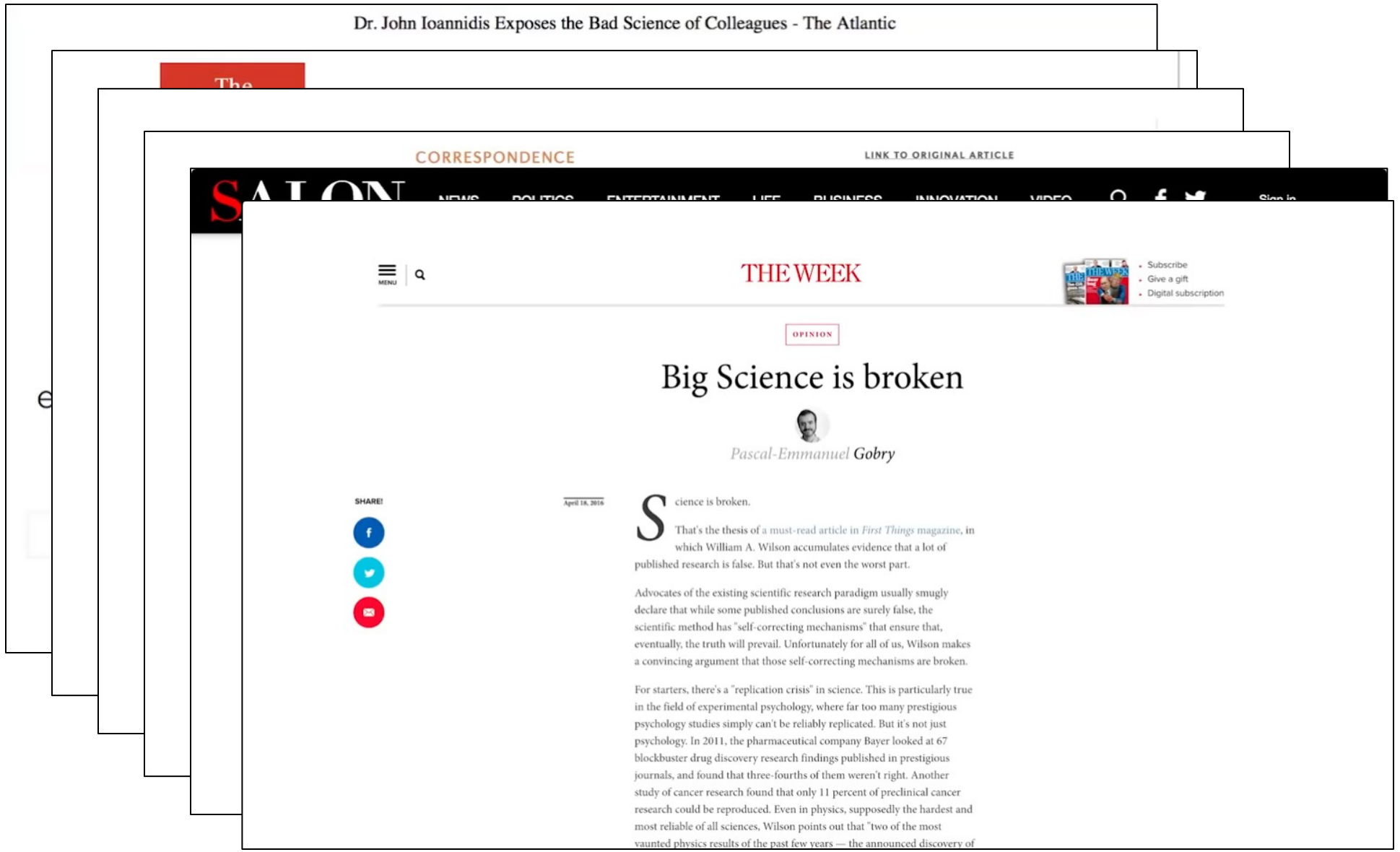
Outline

- The Replication Crisis
- Reproducibility Project: Psychology
- What Does it Mean?
- What Should We Do?

The Replication Crisis

- The Replication Crisis
- Reproducibility Project: Psychology
- What Does it Mean?
- What Should We Do?

The Replication Crisis (Reproducibility Crisis)



[Hen Thom 2017]

The Problem

- **Failure to replicate many published findings, even textbook findings**

- **Research biases**
  - **Publication bias: only significant ($p \leq 0.05$) results published**
  - **Selection bias: only significant results selected for analysis**
  - **Reporting bias: only significant results reported in paper**

- **Replication studies rarely funded, rarely published**
  - **Little incentive to do them**
  - **Therefore, most conducted studies are exploratory in nature**

Evidence

- **Cancer Biology**
  - **2011 Analysis: 95% of cancer drugs fail in clinical trials**
  - **Led to replication studies on drug effectiveness (2011–2012)**
- **In other fields, additional replication studies followed**

Sponsor	%Replicated	Number Replicated
Bayer	21%	14/67
Amgen	11%	6/53
National Institute for Neurological Disorders and Stroke	8%	1/12
ALS Therapy Development Institute	0%	0/47
Reproducibility Project: Psychology	36%	35/97

[Hen Thom 2017]

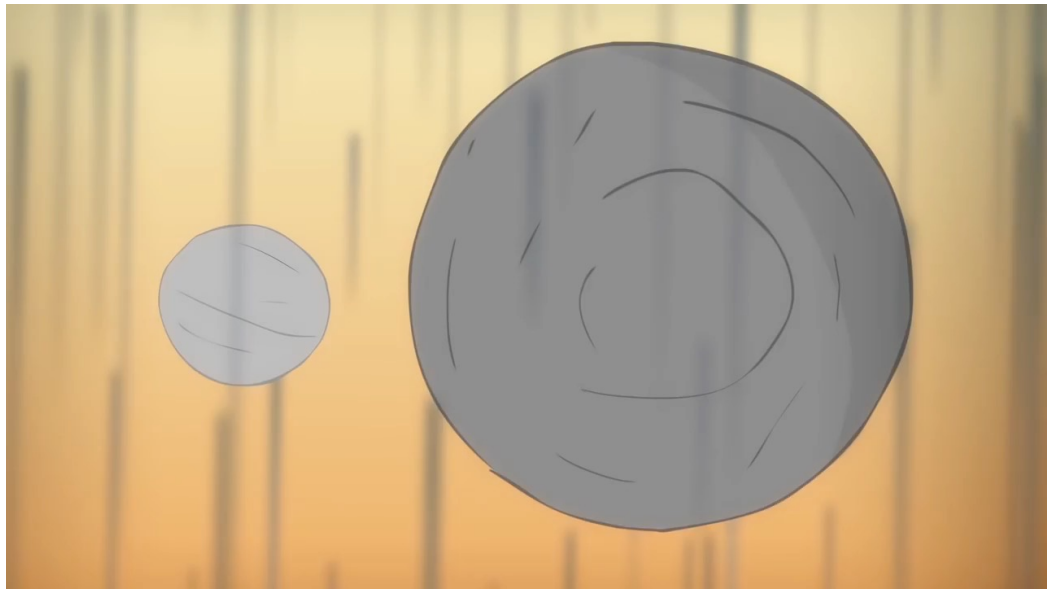
Evidence

- **Replication studies conducted in biomedicine, psychology**
- **Survey data, based on question:**
  - **“Have you failed to reproduce somebody else’s experiment?”**

Field	% Yes
Chemistry	87%
Biology	77%
Physics / Engineering	69%
Medicine	67%
Earth / Environment	64%
Other	62%

[Hen Thom 2017]

The Importance of Replication



[Hen Thom 2017]

Reproducibility Project: Psychology

- The Replication Crisis
- Reproducibility Project: Psychology
- What Does it Mean?
- What Should We Do?

Reproducibility Project: Psychology

Sponsor	%Replicated	Number Replicated
Bayer	21%	14/67
Amgen	11%	6/53
National Institute for Neurological Disorders and Stroke	8%	1/12
ALS Therapy Development Institute	0%	0/47
Reproducibility Project: Psychology	36%	35/97

Reproducibility Project: Psychology

- Begun by Brian Nosek, University of Virginia, 2011
- Replicated 100 published studies
- Recruited very large team
  - Final paper has 270 coauthors
- Which studies to replicate?
  - Goal: minimize selection bias
  - Goal: maximize generalizability
- Published sampling frame and selection criteria

[OSC 2015, 2012]

Sampling frame and selection criteria

- Covered 3 leading journals
  - Psychological Science
  - Journal of Personality and Social Psychology
  - Journal of Experimental Psychology: Learning, Memory, and Cognition

- First 20 articles in each journal, then 10 more; begin with first 2008 issue
- Replicate last study in article (unless infeasible); 84% were last study
- Result must be a single inference test, usually *t*-test, *F*-test, *r* correlation
- If available, use original materials
- Seek design feedback from original authors
- Enough participants for high statistical power (1 – β (power) ≥ 0.80)

Article selection results

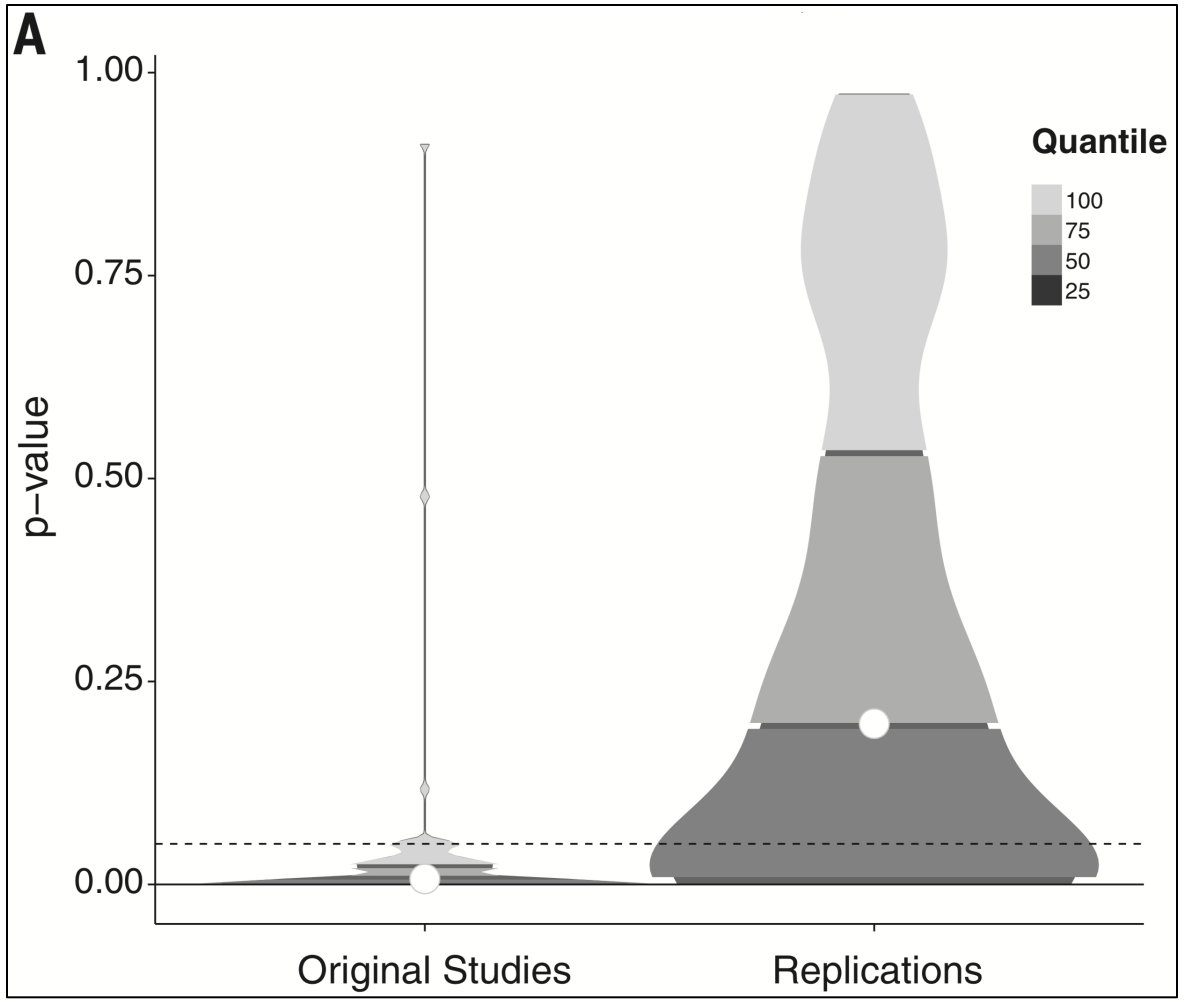
- 488 articles in 2008 issues of the 3 journals
- 158 available for replication
- 113 replications selected
- 100 completed by deadline

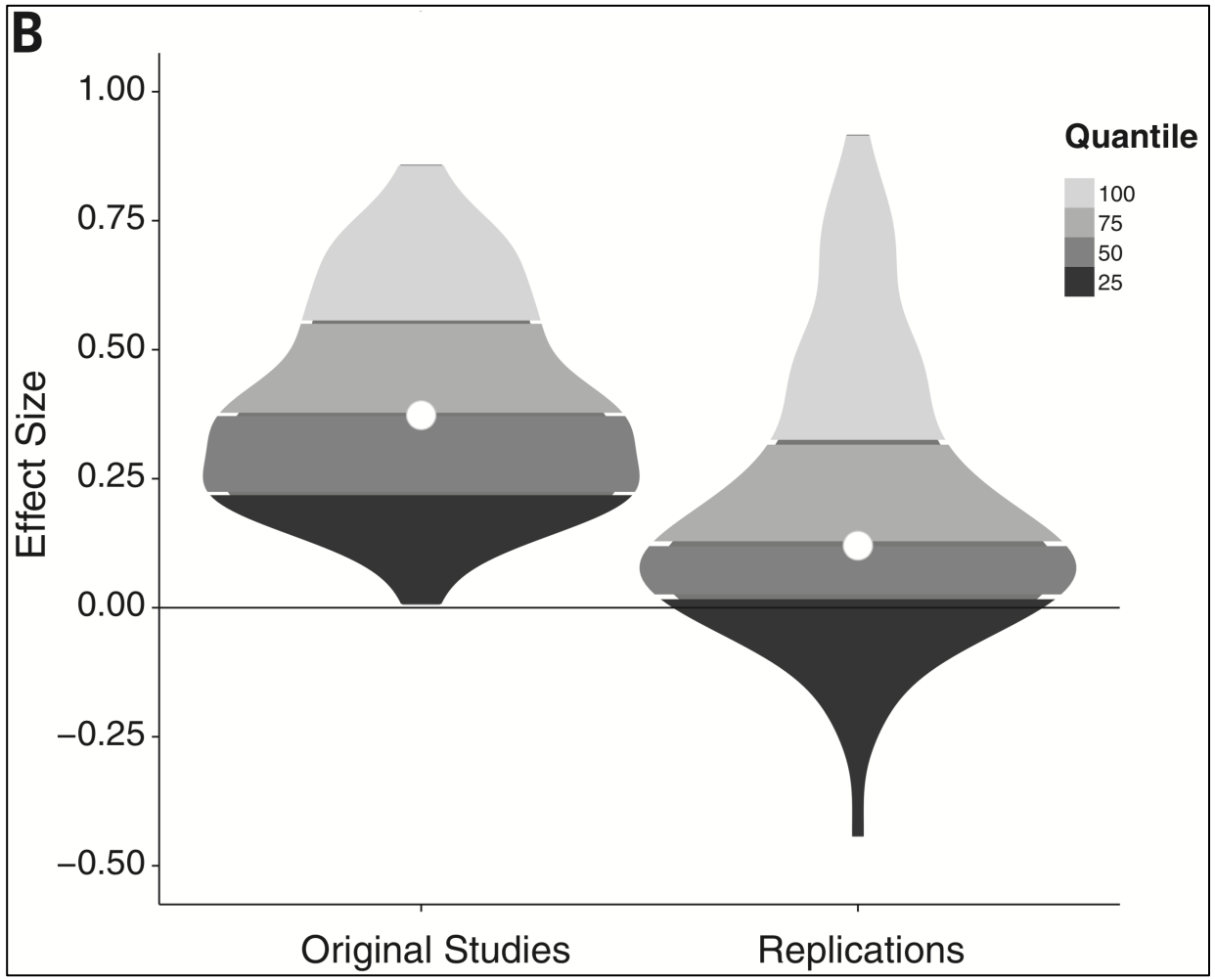
Data collection and processing

- How to measure a replication?
- How to quantify a series of replications?
- Each experiment analyzed with standard R packages
- Each analysis performed independently by 2nd team

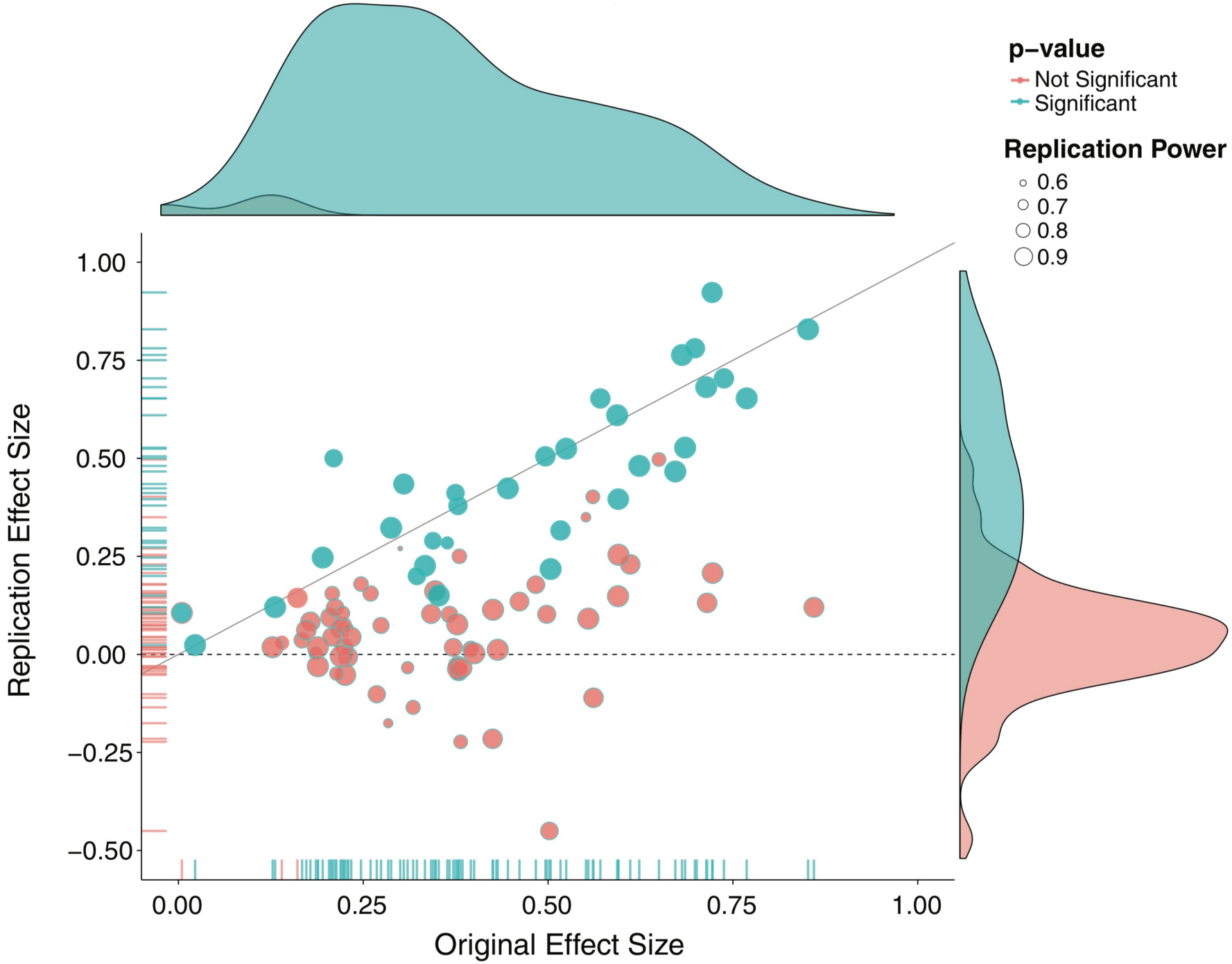
Original Study Result Characteristics	Replication Study Result Characteristics
p value	*p* value
effect size	effect size
df or sample size	*df* or sample size
result importance rating	power
result surprisingness rating	replication challenge rating
experience, expertise rating of original team	experience, expertise rating of replicating team
	replication quality rating

Results


A
1.00
0.75
0.50
0.25
0.00
p–value
Quantile
100
75
50
25
Original Studies
Replications


B
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Effect Size
Quantile
100
75
50
25
Original Studies
Replications

Results


p–value
Not Significant
Significant
Replication Power
0.6
0.7
0.8
0.9
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Replication Effect Size
0.00
0.25
0.50
0.75
1.00
Original Effect Size

Results by %Replicated ($p \leq 0.05$)

- Initial strength of evidence predicts replication success

Original Strength of Evidence	%Replicated ($p \leq 0.05$)	Number Replicated
$p \leq 0.001$	63%	20/32
$p \leq 0.02$	41%	26/63
$0.02 \leq p \leq 0.04$	26%	6/23
$0.04 \leq p$	18%	2/11

- Cognitive psychology more successful than social psychology

Sub-Discipline	%Replicated ($p \leq 0.05$)	Number Replicated
Cognitive Psychology	50%	21/42
Social Psychology	25%	14/55

  - Weaker original effects in social psychology
  - More within-subject, repeated measures designs in cognitive psychology

Results by %Replicated ($p \leq 0.05$)

- Main effects more successful than interactions

Effect Type	%Replicated ($p \leq 0.05$)	Number Replicated
Main Effect	47%	23/49
Interaction Effect	22%	8/37

Results by Correlation with replications ($p \leq 0.05$, original direction)

- Surprising effects were less reproducible ($r = -0.244$)
- Challenging experiments less reproducible ($r = -0.219$)
- Original result importance had little effect ($r = -0.105$)
- Team experience and expertise had almost no effect
  - Original ($r = -0.072$); Replication ($r = -0.096$)
- Replication quality had almost no effect ($r = -0.069$)

- Larger original effect sizes were more reproducible ($r = 0.304$)
- Larger replication effect sizes were more reproducible ($r = 0.731$)
- More powerful replications were more reproducible ($r = 0.368$)

Summary

- **Even though the replications:**
  - **Used materials from original authors**
  - **Were reviewed in advance for methodological fidelity**
  - **Had high statistical power to measure original effect size**

    → replications produced weaker evidence for original findings

- **The strength of initial evidence (*p* value, effect size)**

  → predicted replication success

- **The characteristics of the teams, and the original finding**

  → no impact on replication success

Why so few replications?

- Publication, selection, reporting biases

  → effect sizes of original studies inflated

- Replications
  - All results reported

    → no publication bias
  - All confirmatory tests based on pre-analysis plans

    → no selection, reporting bias

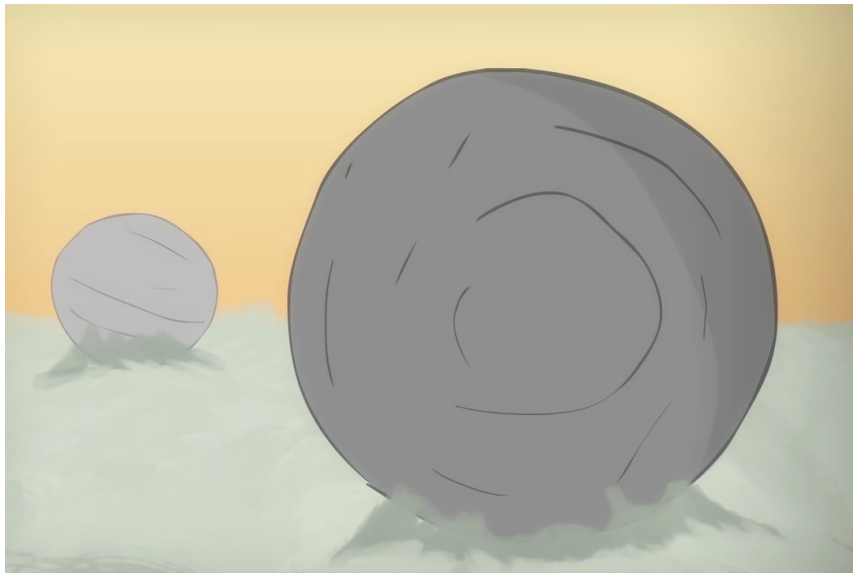
  - Lack of biases likely big part of the reason

What Does it Mean?

- The Replication Crisis
- Reproducibility Project: Psychology
- What Does it Mean?
- What Should We Do?

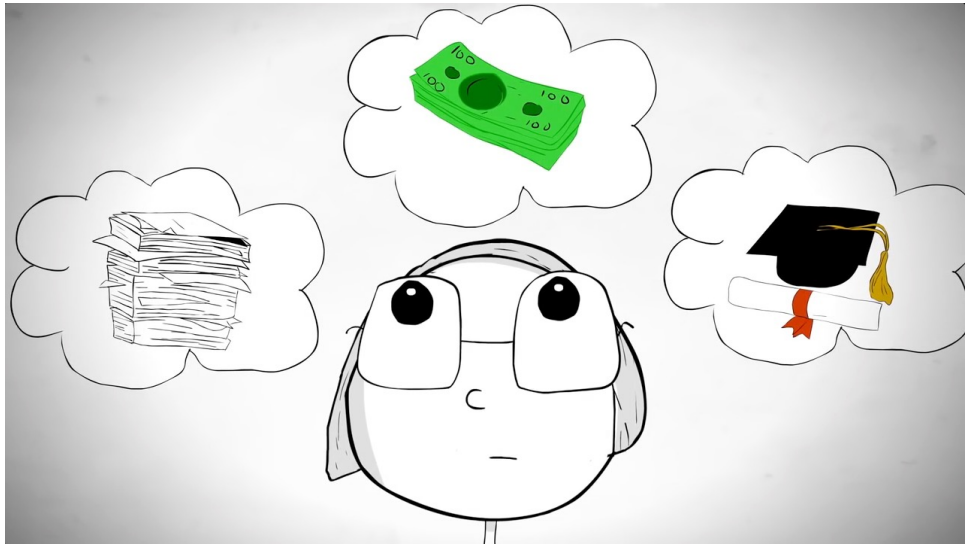
Reasons for Irreproducibility

- A study finds A, but the replication study does not find A. Why?
  1. The original study is wrong → A is not true
  2. The replication study is wrong → A is true
  3. Both original and replication study are correct → A could be true or false
- How could #3 be the case?



Reasons for Irreproducibility

- First impressions are often false
- Can be hard to detect difference between real result and noise
- If enough hypothesis tests are conducted, can usually find something
  - Can be controlled by adjusting familywise α level [Howell 2002, ch 12]
- Incentive structure of science does not maximize yield of true results
  - Incentives result in many exploratory studies
  - True for every field of science
- If a finding is spurious, won't find evidence until replication is attempted



Considering Reproducibility

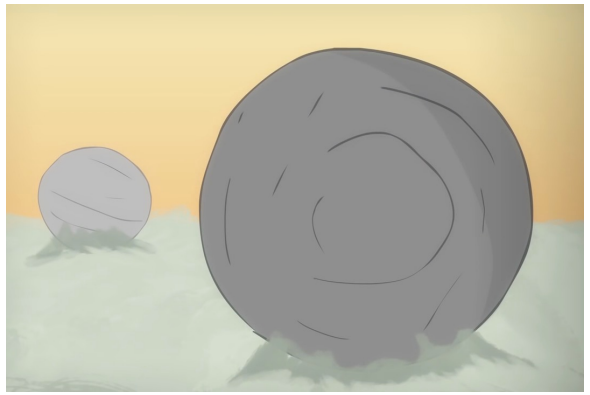
- A study finds A, and the replication study finds A. What does this mean?

  → A is a reliable finding

- What about theoretical explanation for A?

  → Explanation might still be wrong

- Understanding the reasons for A requires multiple investigations
  - Provide converging support for the true theory
  - Rule out alternative, false theories



How Many Studies Should Be Reproducible?

- Is 36% reproducibility too small?
- What would 100% reproducibility mean?
- Progress requires both
  - Exploratory studies: innovative, new ideas
  - Confirmatory studies: replications
- Innovation points out ideas that are possible
- Replication points out ideas that are likely

  → Progress requires both

- Scientific incentives—funding, publication, awards, advancement—should be tuned to encourage an optimal balance, in a collective effort of discovery

What Should We Do?

- The Replication Crisis
- Reproducibility Project: Psychology
- What Does it Mean?
- What Should We Do?

Value (Accept) Replication Studies

- Value confirmation (replication) studies
- Value exploratory studies

  → Value studies that are well done, regardless of type or results

- Requires changing our incentive system

- Less emphasis on surprise
  - "…but rather a reduction in the available cues, which makes the reduced performance not terribly surprising."
  - "…this experiment tells us something important about depth perception in AR, most of which isn't especially surprising, it is not clear that this will help very much…"
  - "It is not entirely surprising that participants became more accurate in 'feedback' condition…"

Recommendations

- Value (accept) replication studies
  - If accepted, they will come
- Pre-register research plans
  - Before collecting data, create detailed, written plan:
    - hypothesis, methods, analysis
  - Removes possibility of p-hacking
  - Even better: publically pre-register the plan
    - e.g., Center for Open Science (https://cos.io)
- Run larger studies
  - more participants == more experimental power
  - BUT: more expensive

Recommendations

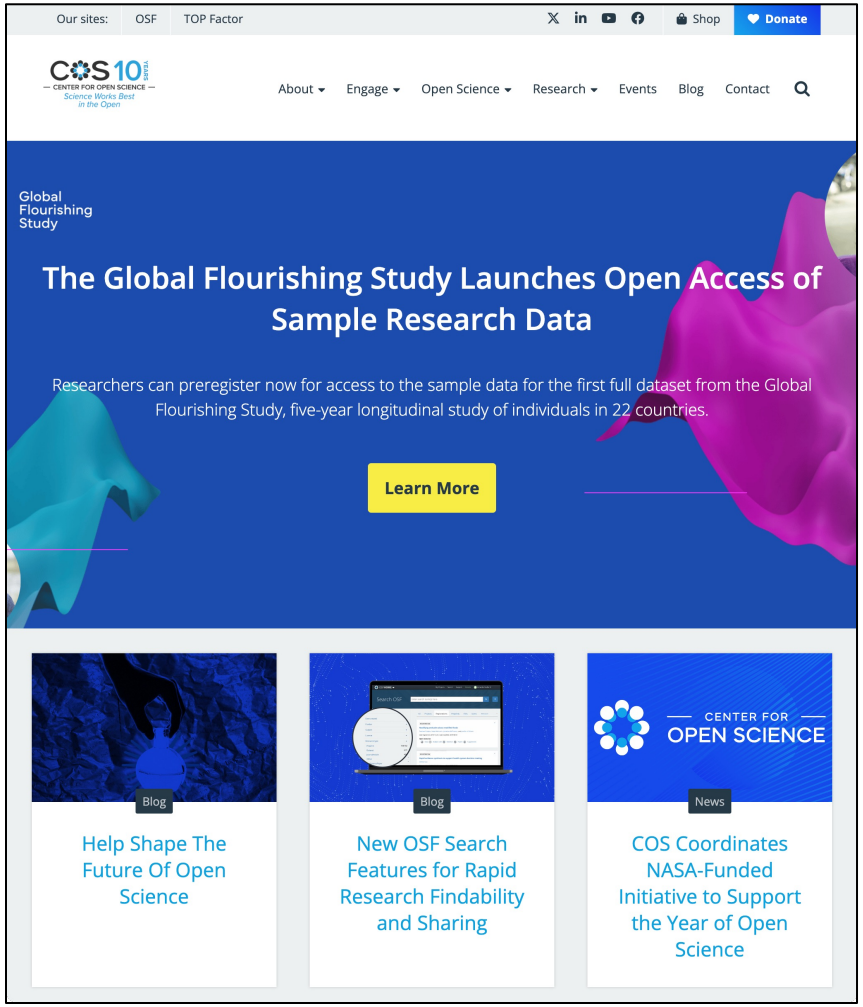
- **Describe methods in more detail → easier replication**
  - **Problem in our field: limited pages**
  - **Solutions:**
    - **Additional details in supplementary material, or in associated thesis / dissertation**
    - **We could adopt longer page limits**
    - **Main paper in bigger font, methods in smaller font (e.g., *Nature*)**
- **Upload materials to open repositories → easier replication**
  - **Data, materials, code**
    - **Center for Open Science (https://cos.io)**
    - **TVCG Replicability Stamp (https://www.computer.org/digital-library/journals/tg/tvcg-replicability-stamp-now-available)**
    - **IEEE DataPort (https://ieee-dataport.org), IEEE Code Ocean (https://codeocean.com)**
    - **arXiv, many other preprint servers, other repositories…**

Conclusion: Reasons for Optimism

- Current zeitgeist among journals, funders, scientists: paying more attention to replication, statistical power, p-hacking, etc.

- In Psychology:
  - Journals have begun publishing pre-registered studies
  - Scientists from many labs have collaboratively replicated earlier studies

- Center for Open Science:
  - Established 2013
  - Developing standards for transparency and openness



References

[Cohen 1994] J Cohen, "The Earth is Round ($p < .05$)", *American Psychologist*, 49(12), pages 997–1003.

[Cohen 1988] J Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition, Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[Economist 2013] "Unreliable Research: Trouble at the Lab", *The Economist*, 18 Oct 2013.

[Freedman 2010] Freedman, D. H., "Lies, Damned Lies, and Medical Science: Dr. John Ioannidis Exposes the Bad Science of Colleagues", *The Atlantic*, Nov 2010.

[Groby 2016] Gobry, P.-E., "Big Science is Broken", *The Week*, 18 April 2016.

[Hen Thom 2017] Henderson, D., Thomson, K., "What Makes Science True?", *NOVA Video Short*, 1 Jan 2017. http://www.pbs.org/wgbh/nova/body/reproduce-science.html

[Ioannidis 2005] Ioannidis, J. P. A., "Why Most Published Research Findings Are False", *PLOS Medicine*, 2(8), e124., 2005. http://doi.org/10.1371/journal.pmed.0020124

[Howell 2002] DC Howell, *Statistical Methods for Psychology*, 5th edition, Duxbury, Pacific Grove, CA, 2002.

[Living Swan et al 2003] MA Livingston, JE Swan II, JL Gabbard, TH Höllerer, D Hix, SJ Julier, Y Baillot, D Brown, "Resolving Multiple Occluded Layers in Augmented Reality", The 2nd *International Symposium on Mixed and Augmented Reality* (ISMAR), 56–65, 2003.

[OSC 2015] Open Science Collaboration, "Estimating the Reproducibility of Psychological Science", *Science*, 349(6251), 2015, DOI: 10.1126/science.aac4716

[OSC 2012] Open Science Collaboration, "An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science", *Perspectives on Psychological Science*, 7(6), 657–660, 2012. http://doi.org/10.1177/1745691612462588

[Prinz et al. 2011] Prinz, F., Schlange, T., & Asadullah, K., "Believe it or not: How much can we rely on published data on potential drug targets?", *Nature Reviews Drug Discovery*, 10(9), 712–712, 2011. http://doi.org/10.1038/nrd3439-c1

[Rehman 2013] Rehman, J., "Cancer research in crisis: Are the drugs we count on based on bad science?", *Salon*, 1 Sep 2013.

[Swan et al 2003] JE Swan II, JL Gabbard, D Hix, RS Schulman, KP Kim, "A Comparative Study of User Performance in a Map-Based Virtual Environment", Technical Papers, *IEEE Virtual Reality*, 259–266, 2003.

[Young 2016] Young, E. (2016, March 4). "Psychology's Replication Crisis Can't Be Wished Away", *The Atlantic*, 4 Mar 2016.

[Young 2015] Young, E., "How Reliable Are Psychology Studies?: Brian Nosek's Reproducibility Project Finds Many Psychology Studies Unreliable", *The Atlantic*, 25 Aug 2015.

Contact Information

J. Edward Swan II

Professor, **Department of Computer Science and Engineering**

Mississippi State University

swan@acm.org

Slide Location:

web.cse.msstate.edu/~swan/teaching/tutorials/Swan-ISMAR2023-WoRXR-Workshop-Replication-XR.pdf

The Replication Crisis, Reproducibility, and the Reproducibility Project in Psychology

J. Edward Swan II

Mississippi State University

Monday, 16 October 2023