Tutorial: October 20, 2018

The Replication Crisis in Empirical Science: Implications for Human Subject Research in Virtual Environments

J. Edward Swan II

Mississippi State University

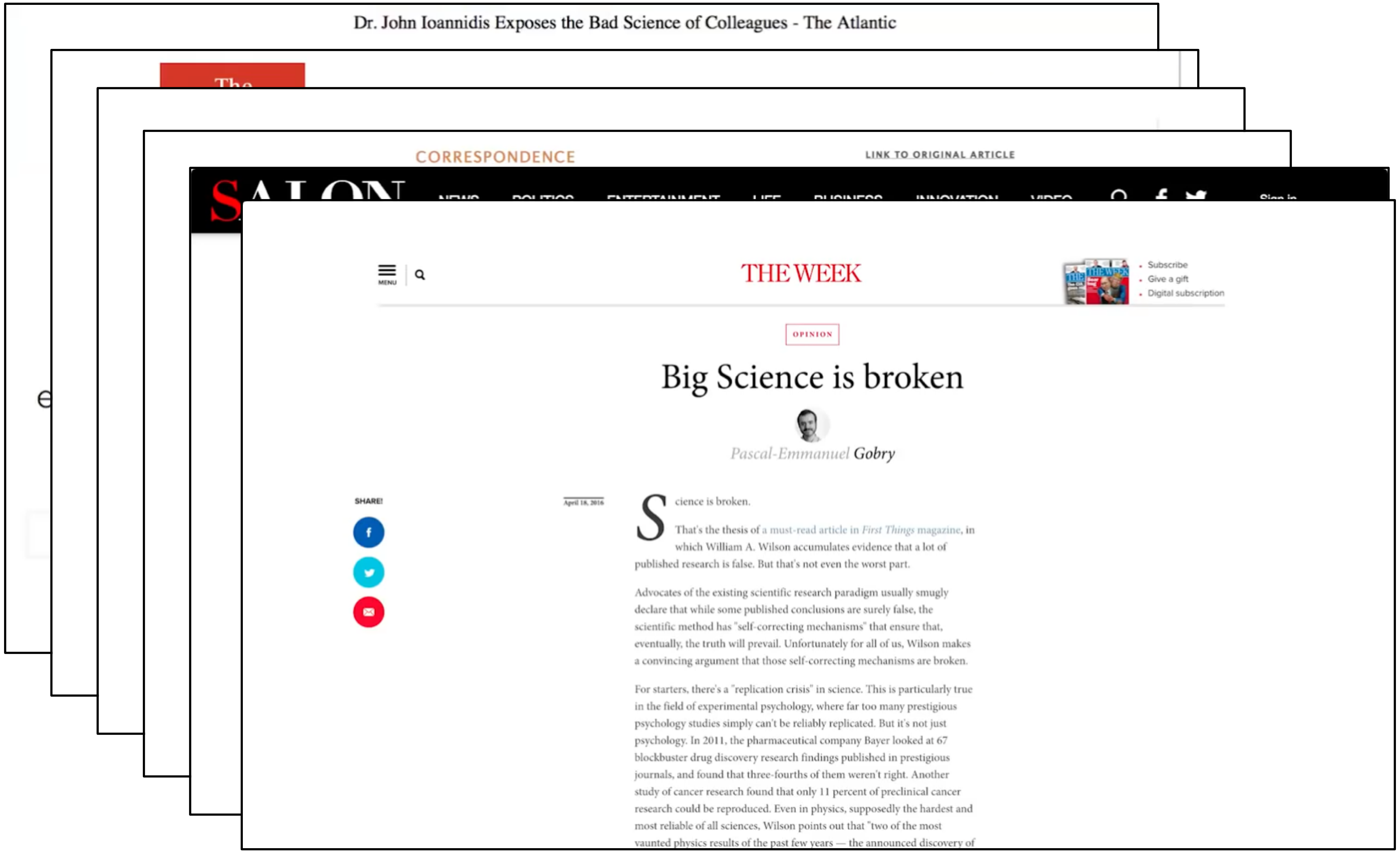
Outline

- The Replication Crisis
- Reproducibility and Inferential Statistics
  - Hypothesis Testing
  - Power, Effect Size, *p*-value
- Reproducibility Project: Psychology
- What Does it Mean?
- What Should We Do?

The Replication Crisis

- The Replication Crisis
- Reproducibility and Inferential Statistics
  - Hypothesis Testing
  - Power, Effect Size, *p*-value
- Reproducibility Project: Psychology
- What Does it Mean?
- What Should We Do?

The Replication Crisis (Reproducibility Crisis)



Dr. John Ioannidis Exposes the Bad Science of Colleagues - The Atlantic

CORRESPONDENCE

LINK TO ORIGINAL ARTICLE

SALON

MENU

THE WEEK

- Subscribe
- Give a gift
- Digital subscription

OPINION

Big Science is broken

Pascal-Emmanuel **Gobry**

SHARE!

April 18, 2016

Science is broken.

That's the thesis of a must-read article in *First Things* magazine, in which William A. Wilson accumulates evidence that a lot of published research is false. But that's not even the worst part.

Advocates of the existing scientific research paradigm usually smugly declare that while some published conclusions are surely false, the scientific method has "self-correcting mechanisms" that ensure that, eventually, the truth will prevail. Unfortunately for all of us, Wilson makes a convincing argument that those self-correcting mechanisms are broken.

For starters, there's a "replication crisis" in science. This is particularly true in the field of experimental psychology, where far too many prestigious psychology studies simply can't be reliably replicated. But it's not just psychology. In 2011, the pharmaceutical company Bayer looked at 67 blockbuster drug discovery research findings published in prestigious journals, and found that three-fourths of them weren't right. Another study of cancer research found that only 11 percent of preclinical cancer research could be reproduced. Even in physics, supposedly the hardest and most reliable of all sciences, Wilson points out that "two of the most vaunted physics results of the past few years — the announced discovery of

[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]

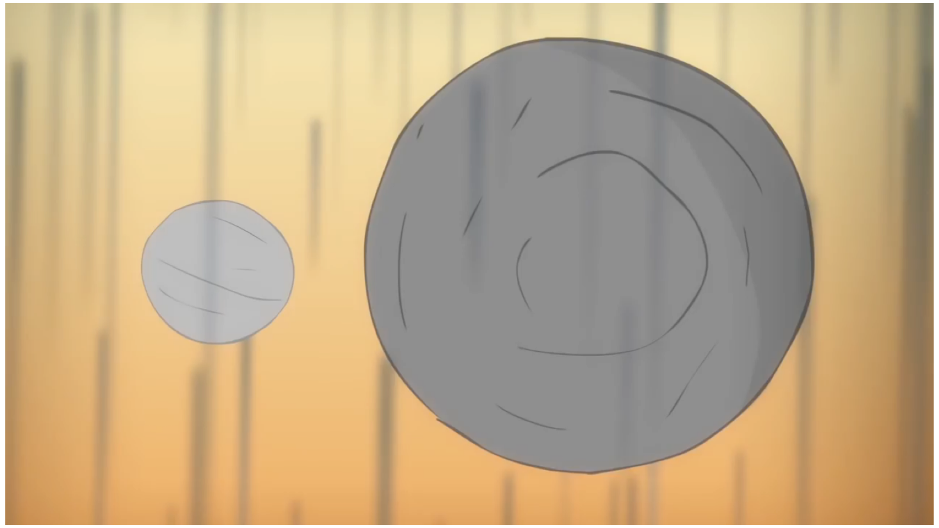
Evidence

- 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



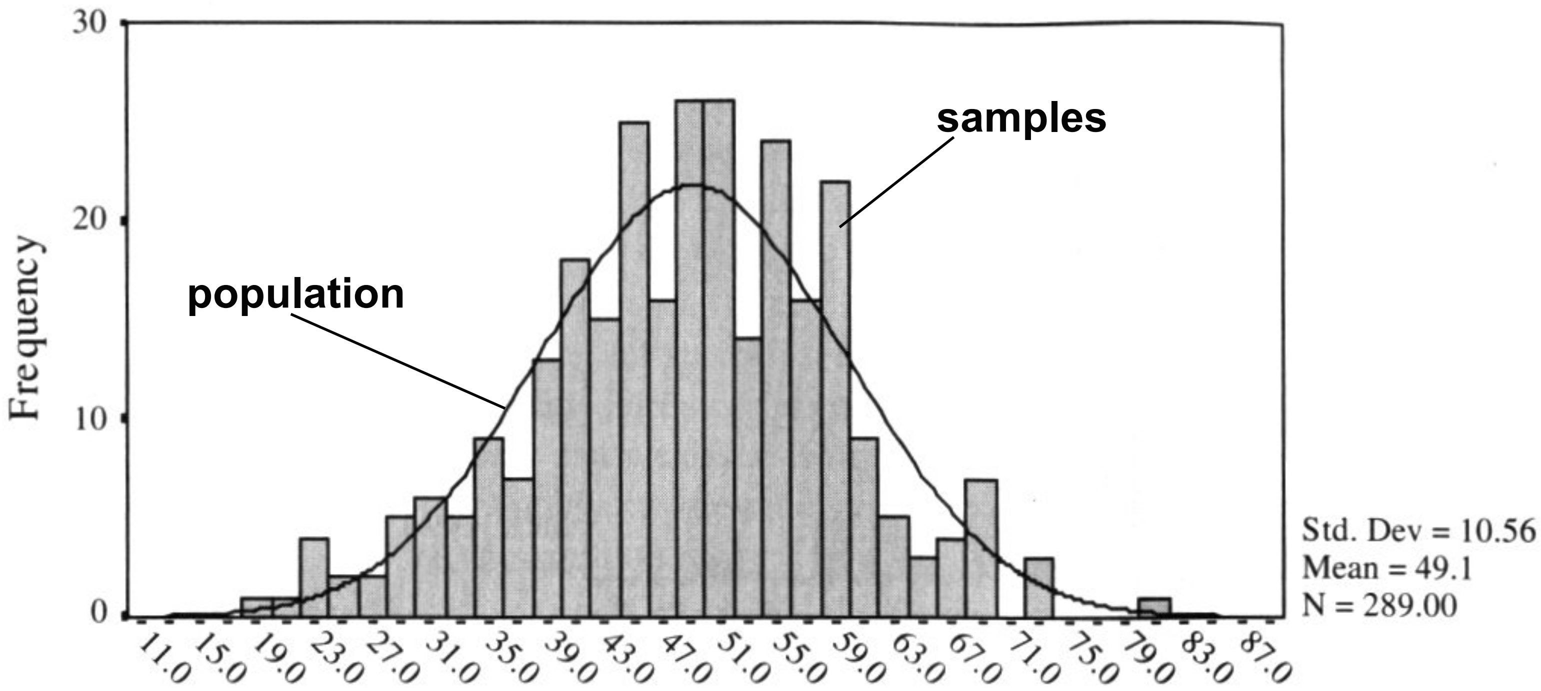
[Hen Thom 2017]

Reproducibility and Inferential Statistics

- The Replication Crisis
- Reproducibility and Inferential Statistics
  - Hypothesis Testing
  - Power, Effect Size, *p*-value
- Reproducibility Project: Psychology
- What Does it Mean?
- What Should We Do?

Hypothesis Testing

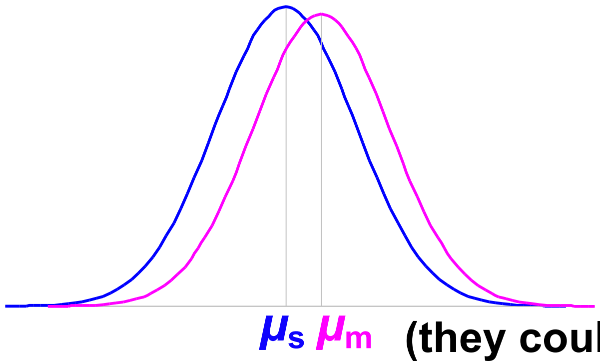
- Goal is to infer population characteristics from sample characteristics



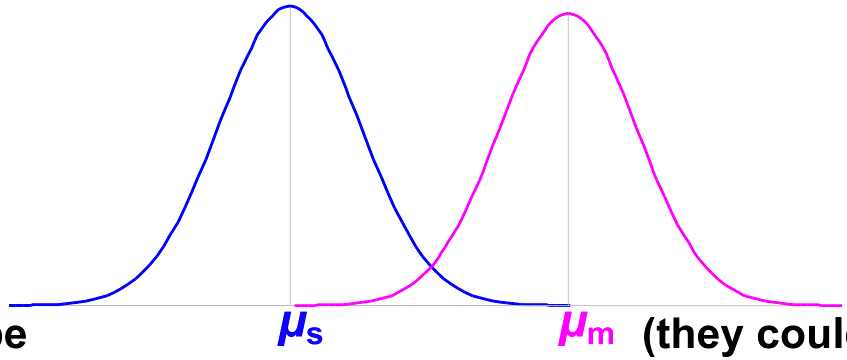
[Howell 2002, p 78]

What Are the Possible Alternatives?

- Let time to navigate be μ_s: stereo time; μ_m: mono time
  - Perhaps there are two populations: $\mu_s - \mu_m = d$

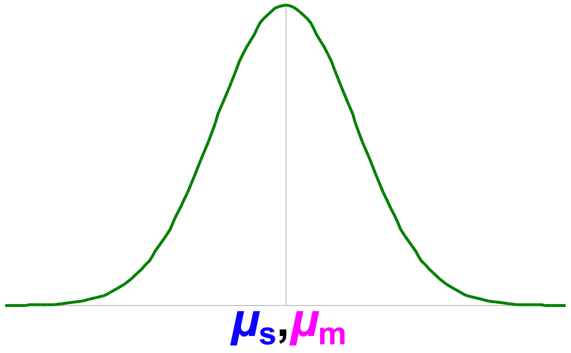


(they could be close together)



(they could be far apart)

  - Perhaps there is one population: $\mu_s - \mu_m = 0$



Hypothesis Testing Procedure

1. Develop testable hypothesis H_1: $\mu_s - \mu_m = d$
   - (E.g., subjects faster under stereo viewing)
2. Develop null hypothesis H_0: $\mu_s - \mu_m = 0$
   - Logical opposite of testable hypothesis
3. Construct sampling distribution assuming H_0 is true.
4. Run an experiment and collect samples; yielding sampling statistic X.
   - (E.g., measure subjects under stereo and mono conditions)
5. Referring to sampling distribution, calculate conditional probability of seeing X given H_0: $p(X \mid H_0)$.
   - If probability is low ($p \leq 0.05$), we are unlikely to see X when H_0 is true. We reject H_0, and embrace H_1.
   - If probability is not low ($p > 0.05$), we are likely to see X when H_0 is true. We do not reject H_0.

Example 1: VE Navigation with Stereo Viewing

1. Hypothesis H_1: $\mu_s - \mu_m = d$
   – Subjects faster under stereo viewing.

2. Null hypothesis H_0: $\mu_s - \mu_m = 0$
   – Subjects same speed whether stereo or mono viewing.

3. Constructed sampling distribution assuming H_0 is true.

4. Ran an experiment and collected samples:
   – 32 participants, collected 128 samples
   – X_s = 36.431 sec; X_m = 34.449 sec; $X_s - X_m$ = 1.983 sec

5. Calculated conditional probability of seeing 1.983 sec given H_0: p(1.983 sec | H_0) = 0.445.
   – p = 0.445 not low, we are likely to see 1.983 sec when H_0 is true. We do not reject H_0.
   – This experiment did not tell us that subjects were faster under stereo viewing.

[Swan et al. 2003]

Example 2: Effect of Intensity on AR Occluded Layer Perception

1. Hypothesis H_1: $\mu_c - \mu_d = d$
   - Tested constant and decreasing intensity. Subjects faster under decreasing intensity.
2. Null hypothesis H_0: $\mu_c - \mu_d = 0$
   - Subjects same speed whether constant or decreasing intensity.
3. Constructed sampling distribution assuming H_0 is true.
4. Ran an experiment and collected samples:
   - 8 participants, collected 1728 samples
   - X_c = 2592.4 msec; X_d = 2339.9 msec; $X_c - X_d$ = 252.5 msec
5. Calculated conditional probability of seeing 252.5 msec given H_0: p(252.5 msec | H_0) = 0.008.
   - $p = 0.008$ is low ($p \leq 0.01$); we are unlikely to see 252.5 msec when H_0 is true. We reject H_0, and embrace H_1.
   - This experiment suggests that subjects are faster under decreasing intensity.

[Living Swan et al. 2003]

Some Considerations…

- The conditional probability $p(X \mid H_0)$
  - Much of statistics involves how to calculate this probability; source of most of statistic's complexity
  - Logic of hypothesis testing the same regardless of how $p(X \mid H_0)$ is calculated
  - If you can calculate $p(X \mid H_0)$, you can test a hypothesis

- The null hypothesis H_0
  - H_0 usually in form $f(\mu_1, \mu_2,\ldots) = 0$
  - Gives hypothesis testing a double-negative logic: assume H_0 as the opposite of H_1, then reject H_0
  - Philosophy is that can never prove $f = 0$, because 0 is point value in domain of real numbers
  - H_1 usually in form $f(\mu_1, \mu_2,\ldots) \neq 0$; we don't know what value it will take, but main interest is that it is not 0

When We Reject H_0

- Calculate $\alpha = p(X \mid H_0)$, when do we reject H_0?
  - In science generally, $\alpha = 0.05$
  - But, just a social convention
- What can we say when we reject H_0 at $\alpha = 0.008$?
  - "If H_0 is true, there is only an 0.008 probability of getting our results, and this is unlikely."
    - Correct!
  - "There is only a 0.008 probability that our result is in error."
    - Wrong, this statement refers to $p(H_0)$, but that's not what we calculated.
  - "There is only a 0.008 probability that H_0 could have been true in this experiment."
    - Wrong, this statement refers to $p(H_0 \mid X)$, but that's not what we calculated.

[Cohen 1994]

When We Don't Reject H_0

- What can we say when we don't reject H_0 at $\alpha = 0.445$?
  - "We have proved that H_0 is true."
  - "Our experiment indicates that H_0 is true."
    - Wrong, hypothesis testing cannot prove H_0: $f(\mu_1, \mu_2, \ldots) = 0$.

- Statisticians do not agree on what failing to reject H_0 means.
  - Conservative viewpoint (Fisher):
    - We must suspend judgment, and cannot say anything about the truth of H_0.
  - Alternative viewpoint (Neyman & Pearson):
    - We can accept H_0 if we have sufficient experimental power, and therefore a low probability of type II error.

[Howell 2002, p 99]

Probabilistic Reasoning

- If hypothesis testing was absolute:
  - If H_0 is true, then X cannot occur…however, X has occurred…therefore H_0 is false.
  - e.g.: If a person is a Martian, then they are not a member of Congress (true)…this person is a member of Congress…therefore they are not a Martian. (correct result)
  - e.g.: If a person is an American, then they are not a member of Congress (false)…this person is a member of Congress…therefore they are not an American. (incorrect result, but correct logical reasoning)

p	q	$p \rightarrow q$	$\neg q \rightarrow \neg p$
T	T	T	T
T	F	F	F
F	T	T	T
F	F	T	T

$$\begin{array}{l} p \rightarrow q \\ \neg q \\ \hline \rightarrow \neg p \end{array}$$

modus tollens

[Cohen 1994]

Probabilistic Reasoning

- However, hypothesis testing is probabilistic:
  - If H_0 is true, then X is highly unlikely…however, X has occurred…therefore H_0 is highly unlikely.
  - e.g.: If a person is an American, then they are probably not a member of Congress (true, right?)…this person is a member of Congress…therefore they are probably not an American.
    (incorrect result, but correct hypothesis testing reasoning)

p	q	$p \to q$	$\neg q \to \neg p$
T	T	T	T
T	F	F	F
F	T	T	T
F	F	T	T

$$\begin{array}{l} p \to q \\ \neg q \\ \hline \to \neg p \end{array} \Bigg\} \text{ modus tollens}$$

[Cohen 1994]

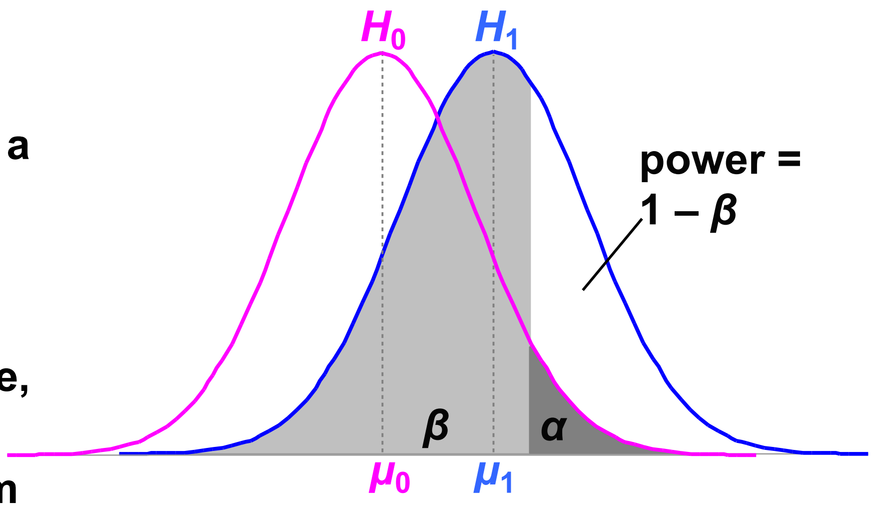
Reproducibility and Inferential Statistics

- The Replication Crisis
- Reproducibility and Inferential Statistics
  - Hypothesis Testing
  - Power, Effect Size, *p*-value
- Reproducibility Project: Psychology
- What Does it Mean?
- What Should We Do?

Interpreting α, β, and Power

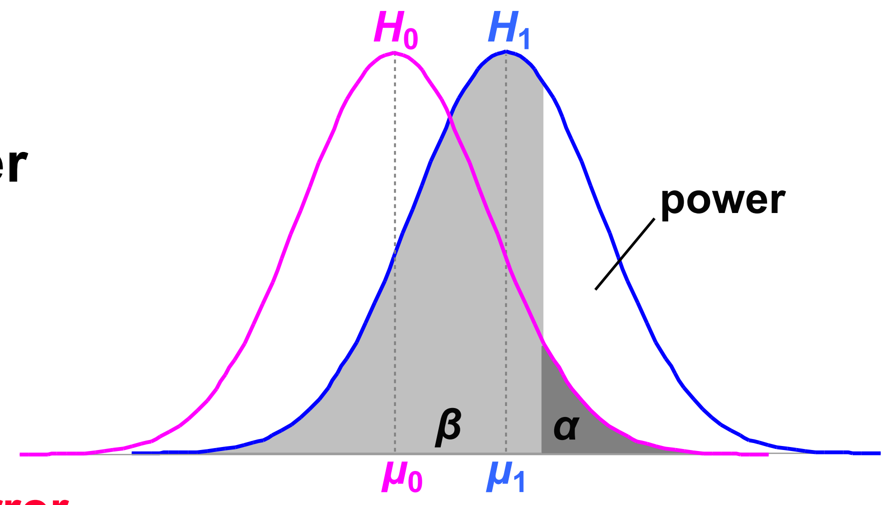
		Decision	
		Reject H_0	Don't reject H_0
True state of the world	H_0 false	a result! $p = 1 - \beta$ = power	type II error $p = \beta$
	H_0 true	type I error $p = \alpha$	argue H_0? $p = 1 - \alpha$

- If H_0 is true:
  - α is probability we make a type I error: we think we have a result, but we are wrong
- If H_1 is true:
  - β is probability we make a type II error: a result was there, but we missed it
  - Power is a more common term than β

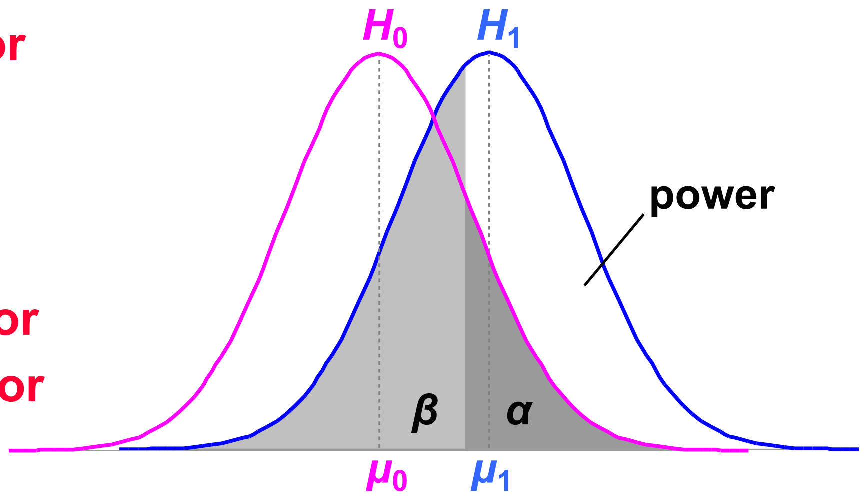


Increasing Power by Increasing α

- Illustrates α / power tradeoff



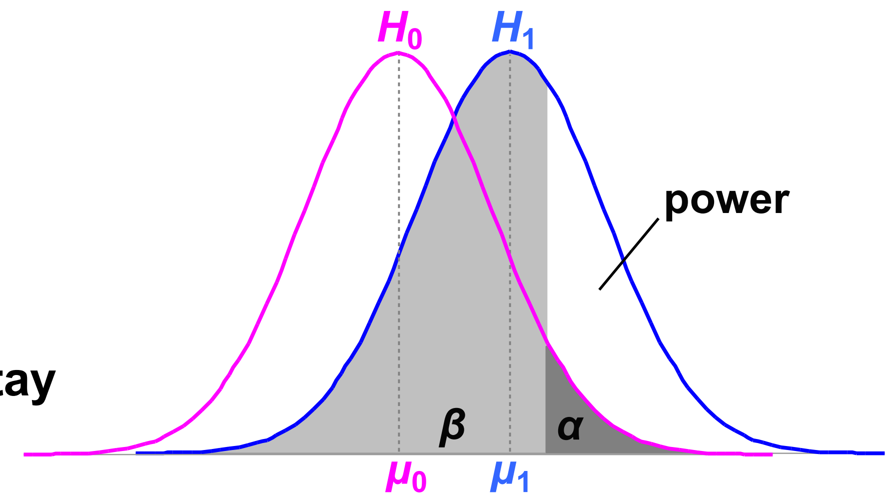
- Increasing α:
  - Increases power
  - Decreases type II error
  - Increases type I error



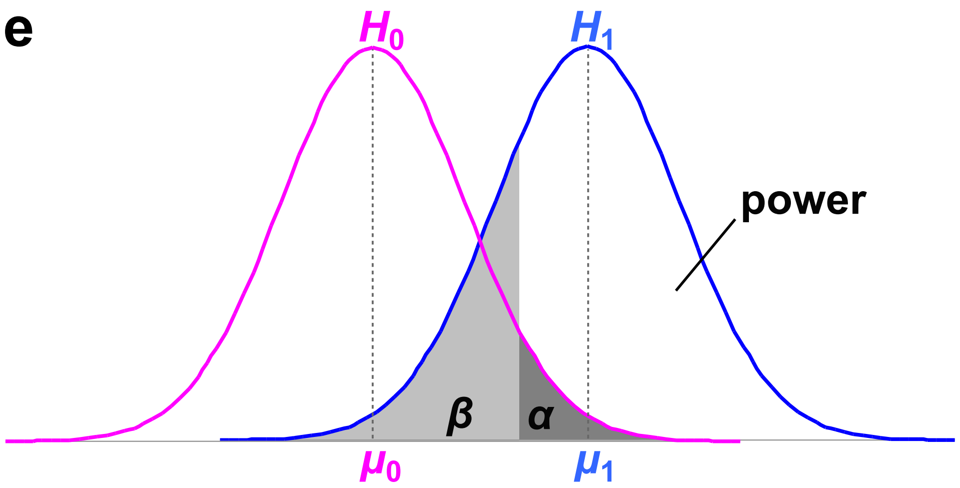
- Decreasing α:
  - Decreases power
  - Increases type II error
  - Decreases type I error

Increasing Power by Measuring a Bigger Effect

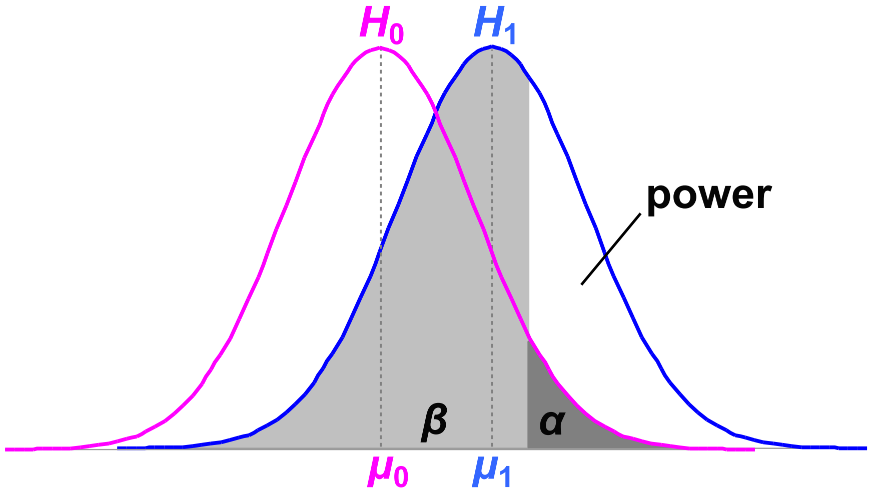
- If the effect size is large:
  - Power increases
  - Type II error decreases
  - α and type I error stay the same



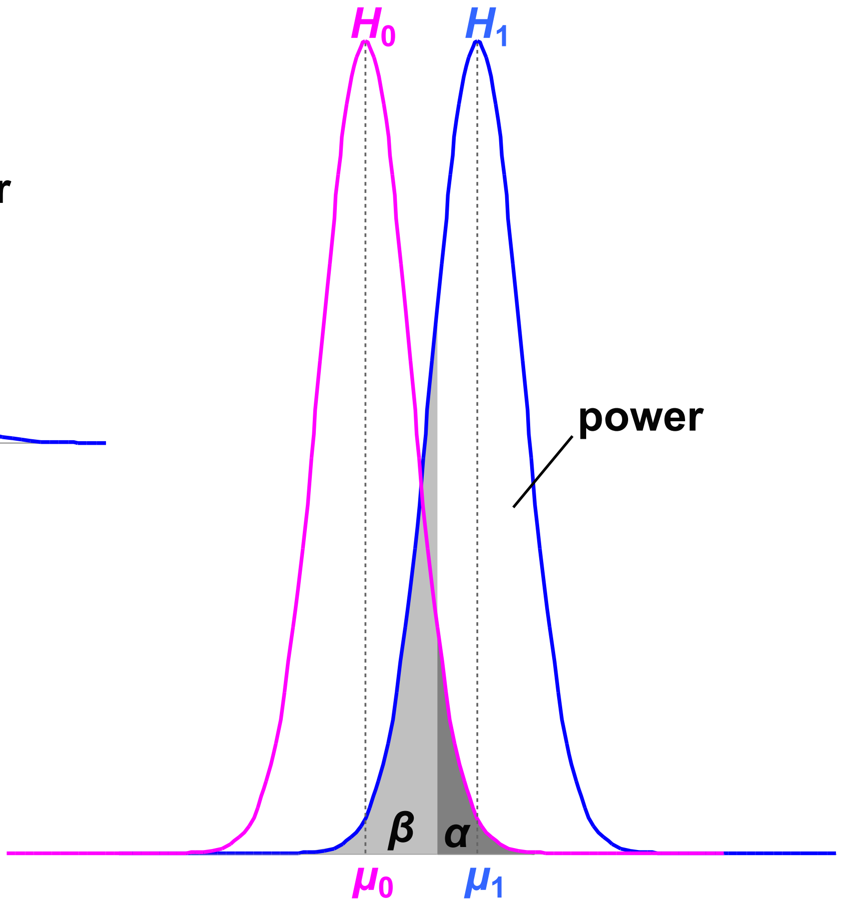
- Unsurprisingly, large effects are easier to detect than small effects



Increasing Power by Collecting More Data

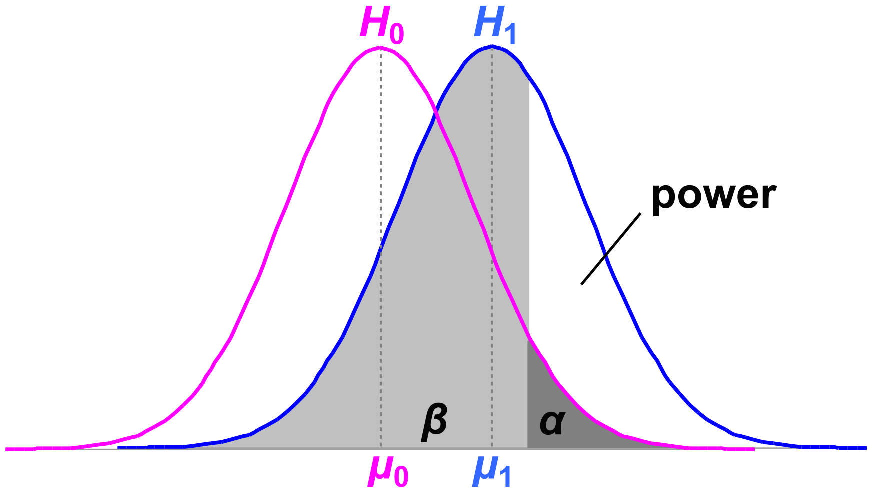


- Increasing sample size (*N*):
  - Decreases variance
  - Increases power
  - Decreases type II error
  - *α* and type I error stay the same
- There are techniques that give the value of *N* required for a certain power level.

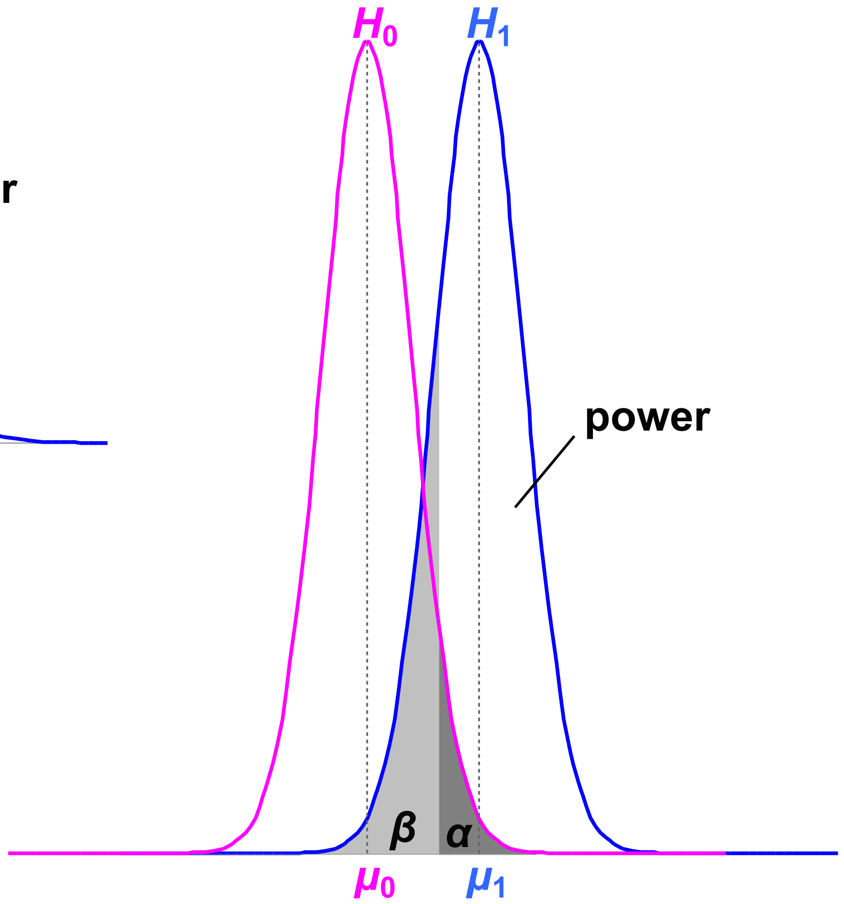


- Here, effect size remains the same, but variance drops by half.

Increasing Power by Decreasing Noise



- **Decreasing experimental noise:**
  - Decreases variance
  - Increases power
  - Decreases type II error
  - α and type I error stay the same
- **More careful experimental results give lower noise.**



- **Here, effect size remains the same, but variance drops by half.**

Using Power

- **Need α, effect size, and sample size for power:**

$$\text{power} = f(\ \alpha,\ |\mu_0 - \mu_1|,\ N\ )$$

- **Problem for VR / AR:**
  - **Effect size $|\mu_0 - \mu_1|$ hard to know in our field**
    - **Population parameters estimated from prior studies**
    - **But our field is so new, not many prior studies**
  - **Can find effect sizes in more mature fields**

- **Post-hoc power analysis:**

$$\text{effect size} = |\bar{X}_0 - \bar{X}_1|$$

  - **Then, calculate power for experiment**
  - **But this makes statisticians grumble (e.g. [Howell 2002] [Cohen 1988])**
  - **Same information as *p* value**

Other Uses for Power

1. Number samples needed for certain power level:

$$N = f(\text{ power}, \alpha, |\mu_0 - \mu_1| \text{ or } |X_0 - X_1| \text{ })$$

   - Number extra samples needed for more powerful result
   - Gives "rational basis" for deciding *N*
   - Cohen [1988] recommends $\alpha = 0.05$, power = 0.80

2. Effect size that will be detectable:

$$|\mu_0 - \mu_1| = f(\text{ } N, \text{power}, \alpha \text{ })$$

3. Significance level needed:

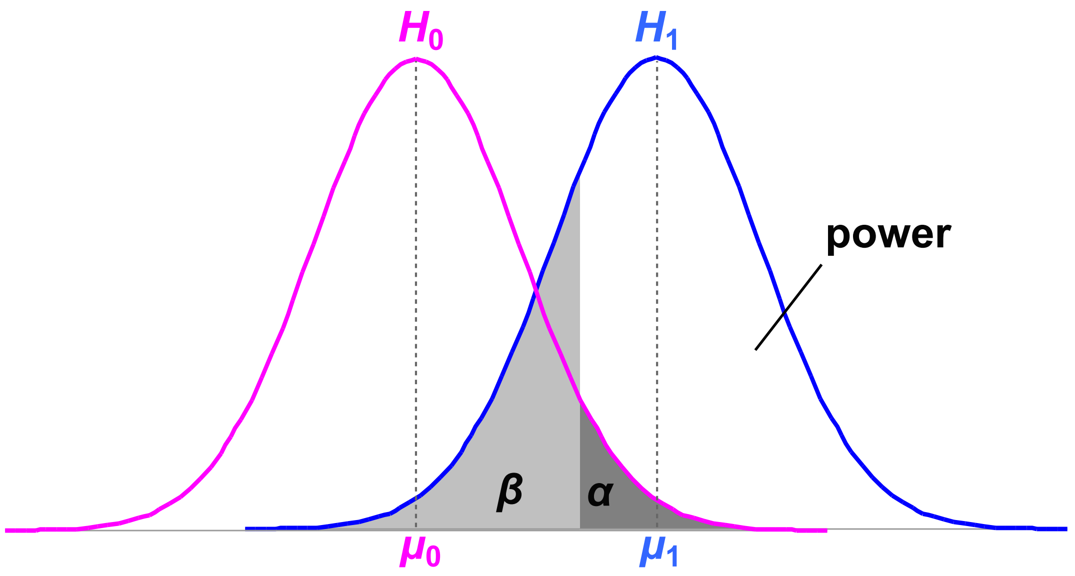
$$\alpha = f(\text{ } |\mu_0 - \mu_1| \text{ or } |X_0 - X_1|, N, \text{power })$$

(1) is the most common power usage

[Cohen 1988]

Arguing the Null Hypothesis

- Cannot directly argue H_0: $\mu_s - \mu_m = 0$. But we can argue that $|\mu_0 - \mu_1| < d$.
  - Thus, we have bound our effect size by *d*.
  - If *d* is *small*, effectively argued null hypothesis.
  - Cohen recommends $\alpha = 0.05$, power = 0.20



[Cohen 1988, p 16]

Reproducibility Project: Psychology

- The Replication Crisis
- Reproducibility and Inferential Statistics
  - Hypothesis Testing
  - Power, Effect Size, *p*-value
- 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 ($1 - \beta$ (power) ≥ 0.80)

Article selection results

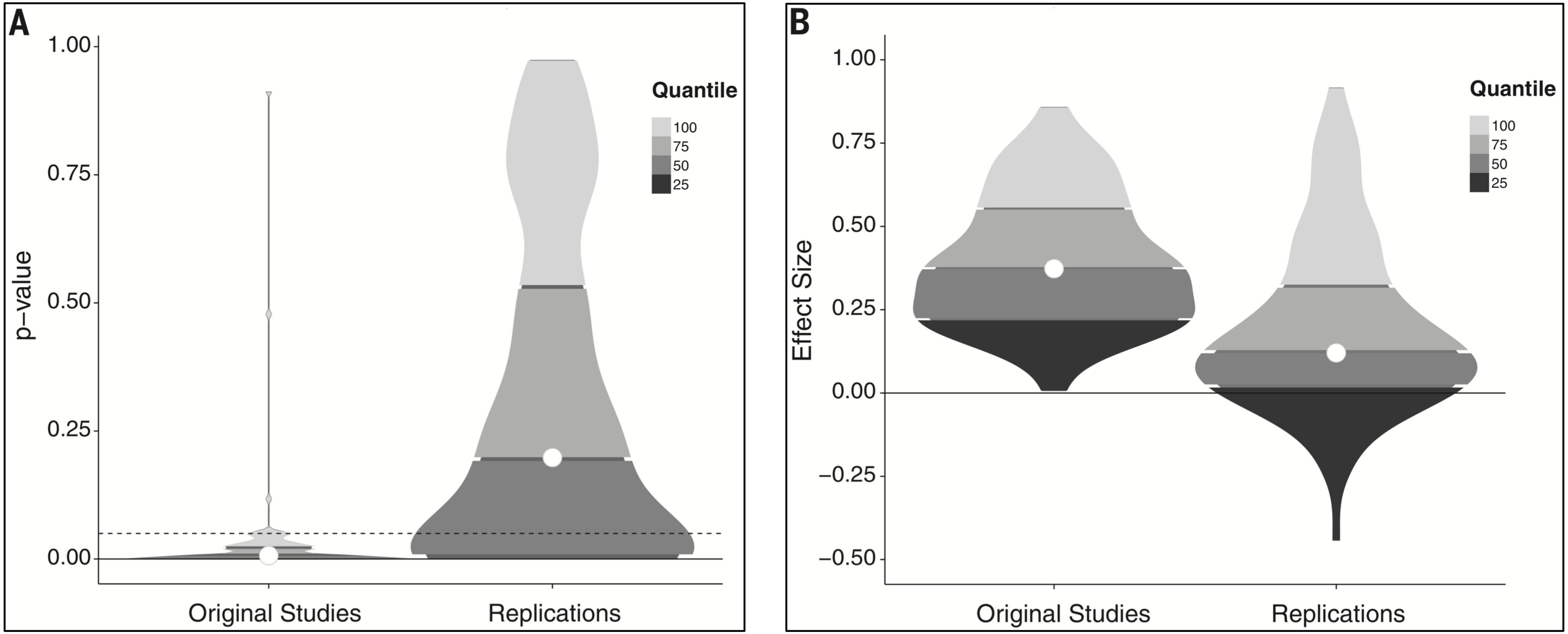
- 488 articles in 2008 issues of the 3 journals
- 158 available for replication
- 113 replications selected
- 100 completed by deadline

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

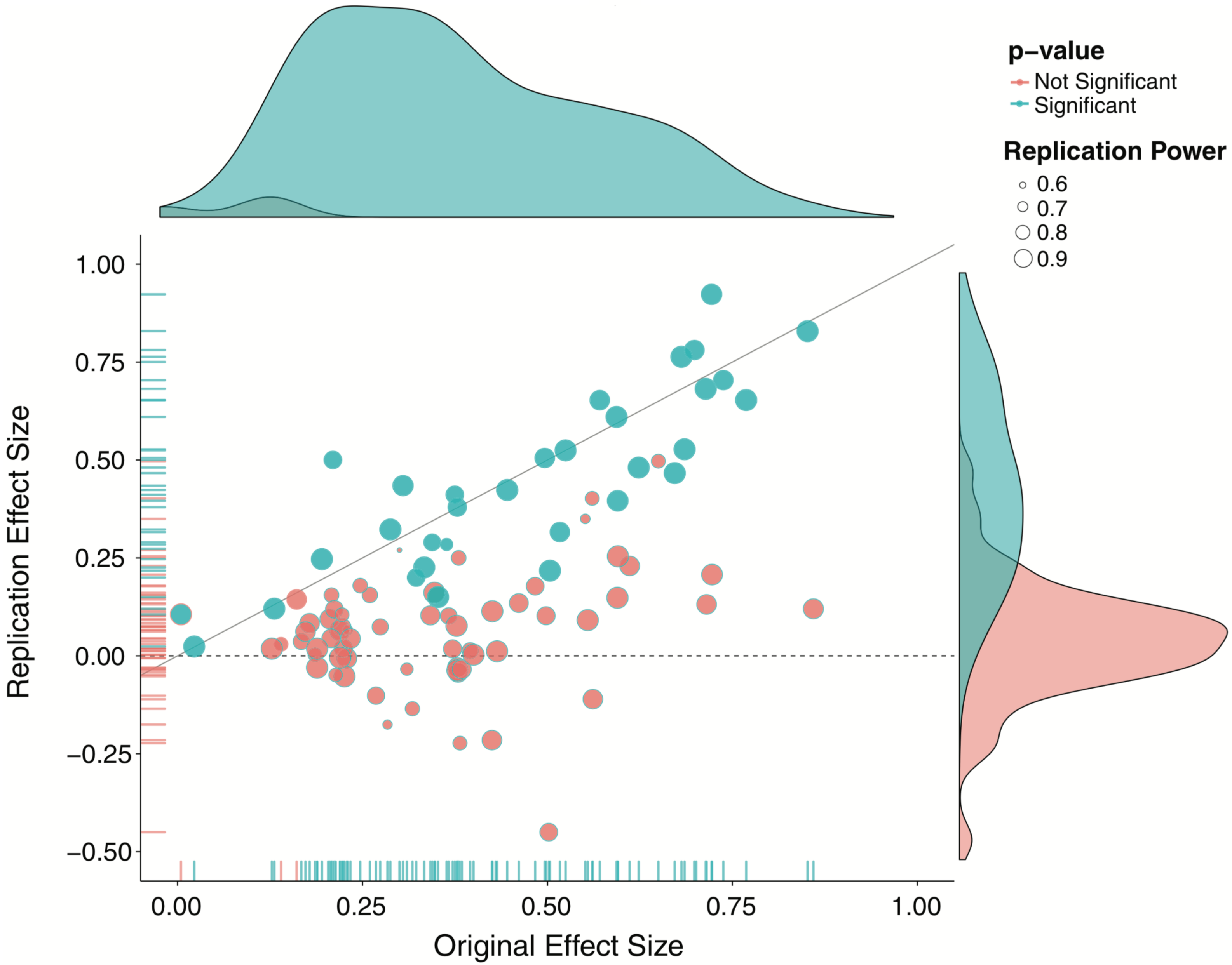
Original Study Result Characteristics	Replication Study Result Characteristics
p value	*p* value
effect size	effect size
df or sample size	*df* or sample size
result importance rating	power
result surprisingness rating	replication challenge rating
experience, expertise rating of original team	experience, 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
Replication Effect Size
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
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 ($r = 0.731$)

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

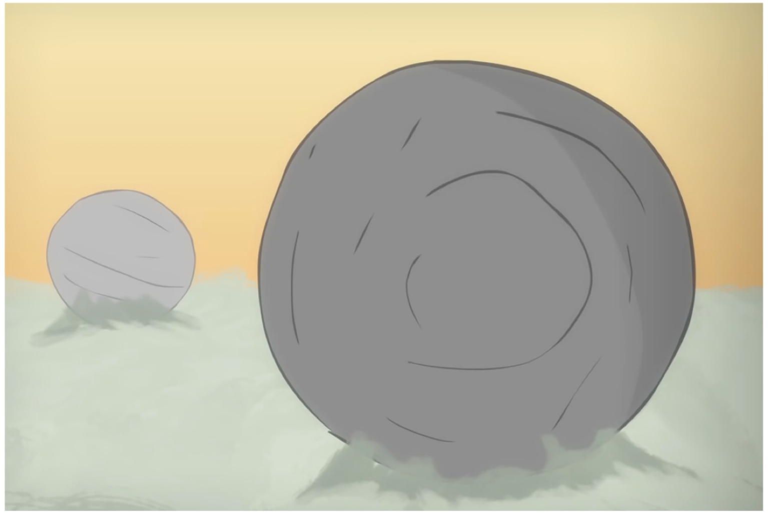
  - Lack of biases likely big part of the reason

What Does it Mean?

- The Replication Crisis
- Reproducibility and Inferential Statistics
  - Hypothesis Testing
  - Power, Effect Size, *p*-value
- Reproducibility Project: Psychology
- What Does it Mean?
- What Should We Do?

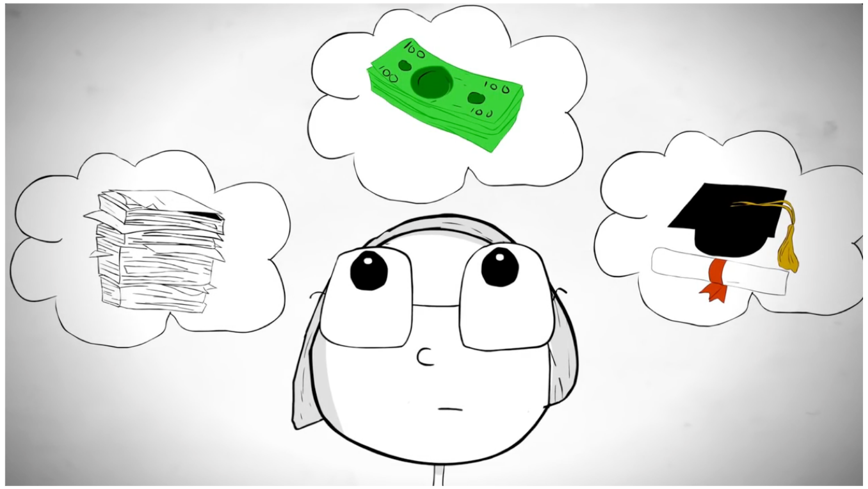
Reasons for Irreproducibility

- 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

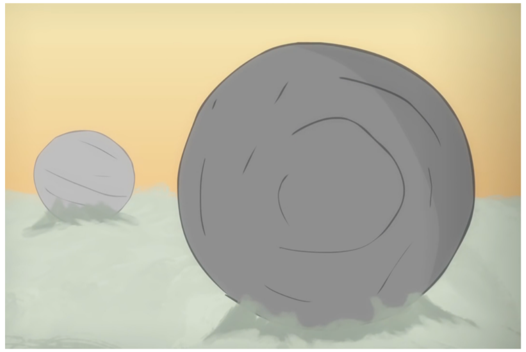
- A study finds A, and the replication study finds A. What does this mean?

  → 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 and Inferential Statistics
  - Hypothesis Testing
  - Power, Effect Size, *p*-value
- 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

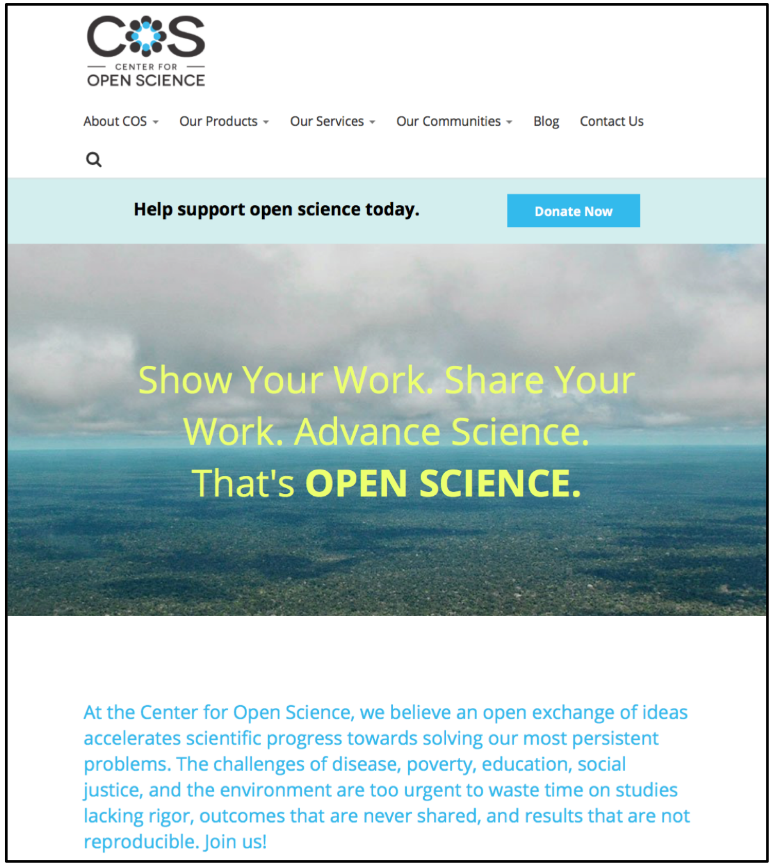
- Value (accept) replication studies
  - If accepted, they will come
- Pre-register research plans
  - Before collecting data, create detailed, written plan:
    - hypothesis, methods, analysis
  - Removes possibility of p-hacking
  - Even better: publically pre-register the plan
    - e.g., Center for Open Science (https://cos.io) → Preregistration Challenge (https://cos.io/prereg/)
- 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)
    - 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
  - Channeling 1M USD to pre-registration challenge



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

Faculty, Center for Advanced Vehicular Systems

Mississippi State University

swan@acm.org

+1-662-325-7507

Slide Location:

http://web.cse.msstate.edu/~swan/teaching/tutorials/Swan-ISMAR2018-Replication-Crisis.pdf

The Replication Crisis in Empirical Science: Implications for Human Subject Research in Virtual Environments

J. Edward Swan II

Mississippi State University