The Replication Crisis in Empirical Science: Implications for Human Subject Research in Mixed Reality

J. Edward Swan II, Mohammed Safayet Arefin

Mississippi State University

Sunday, 28 March 2021

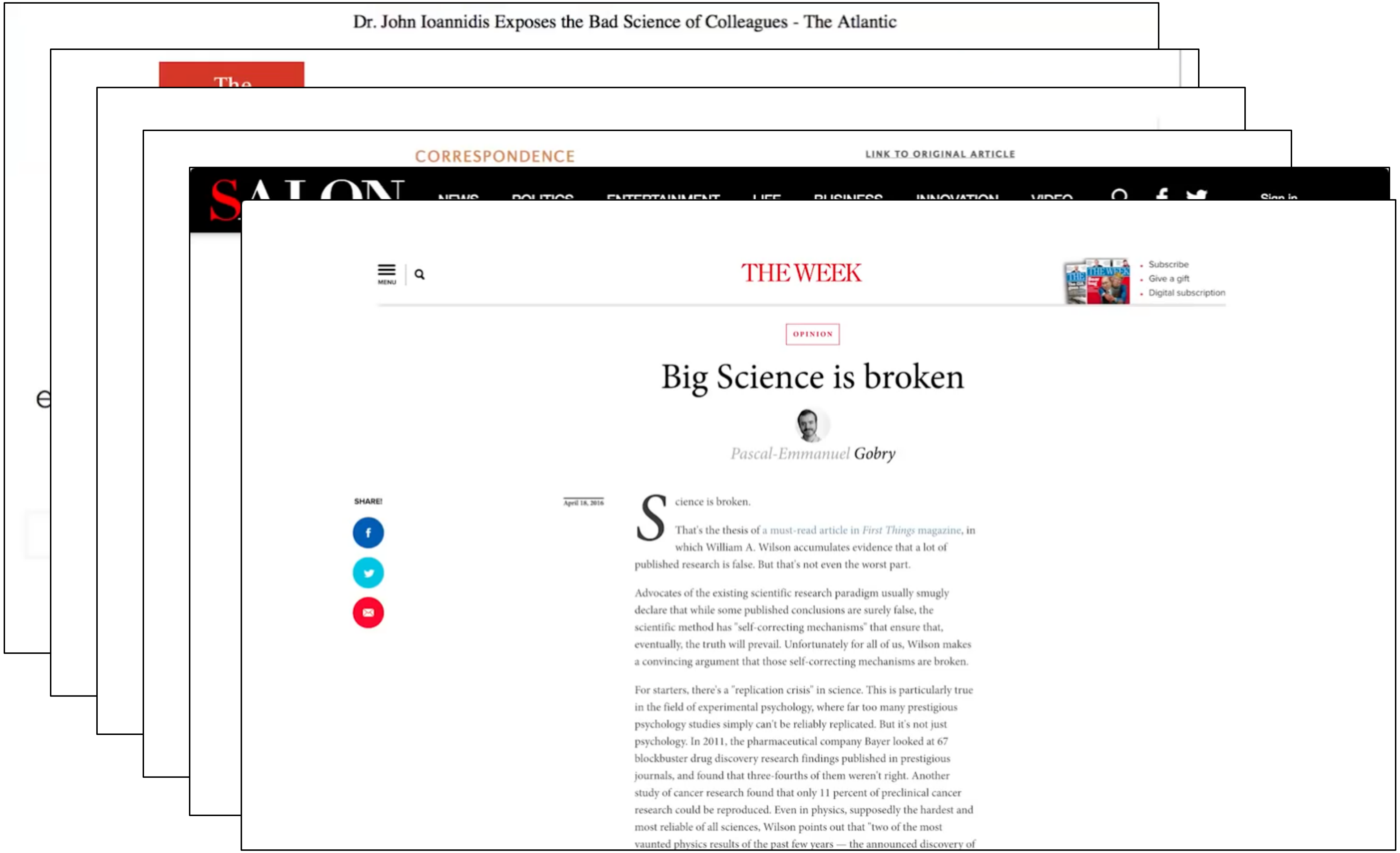
Outline

- 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

THE WEEK

OPINION

Big Science is broken

Pascal-Emmanuel Gobry

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]

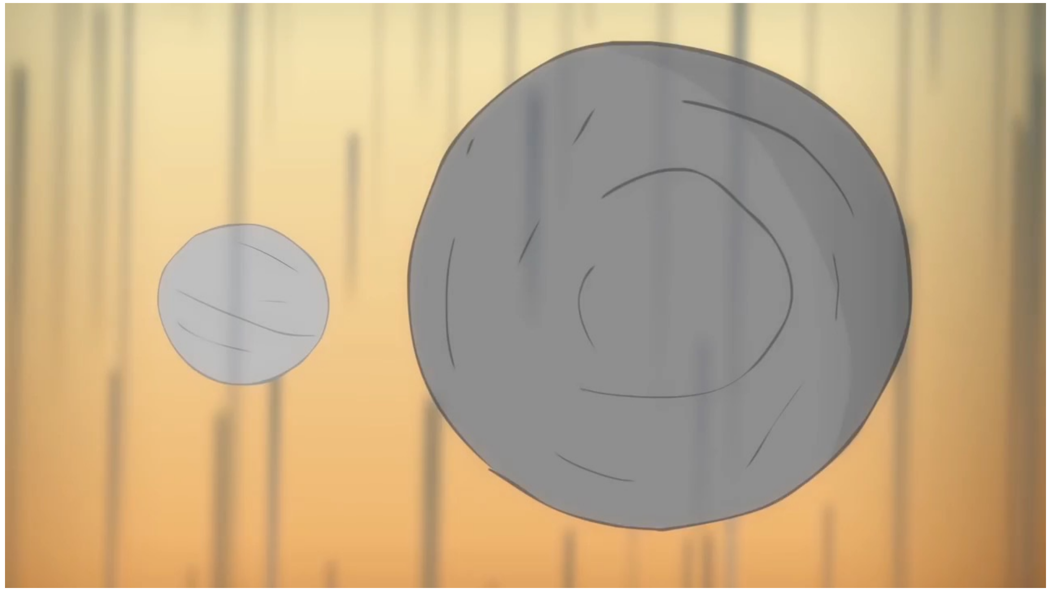
Evidence

- 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



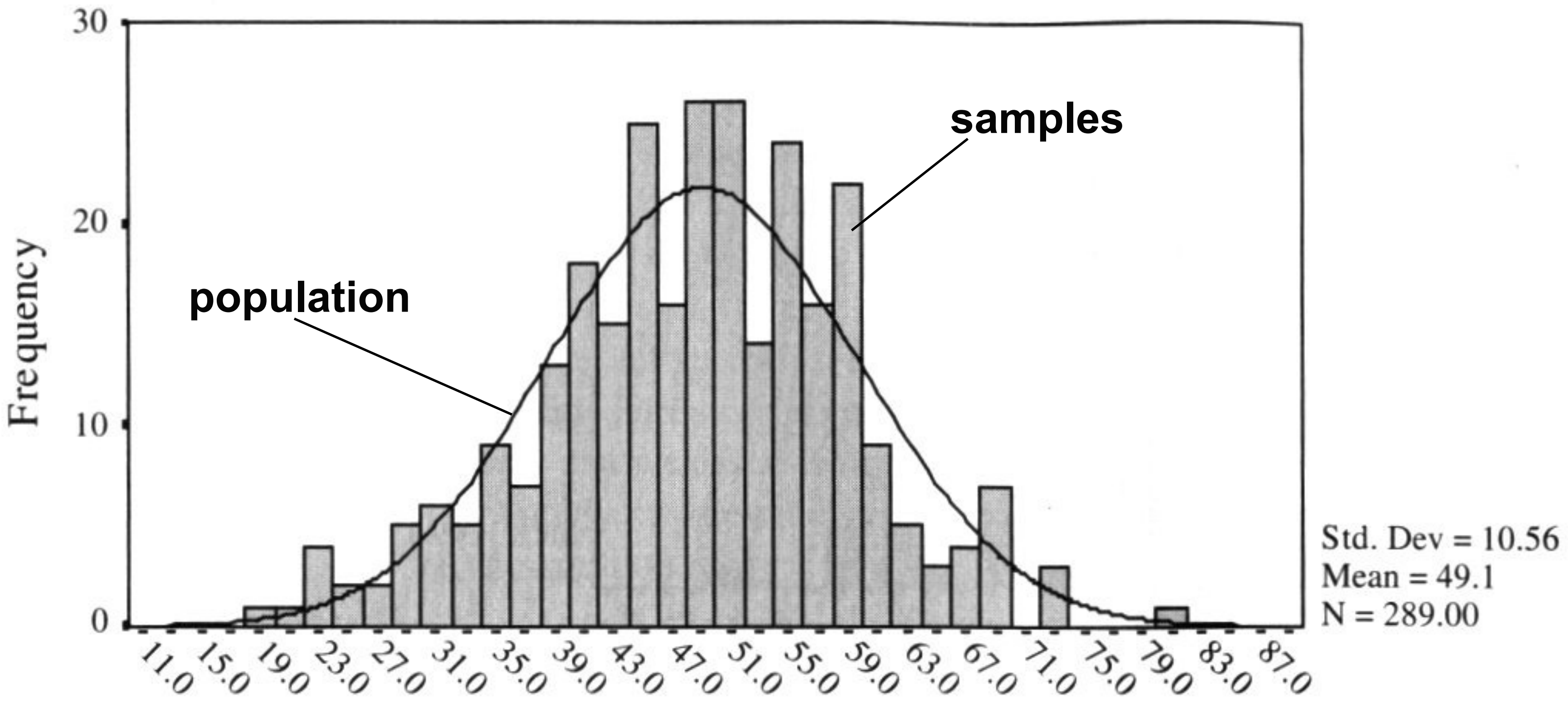
[Hen Thom 2017]

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

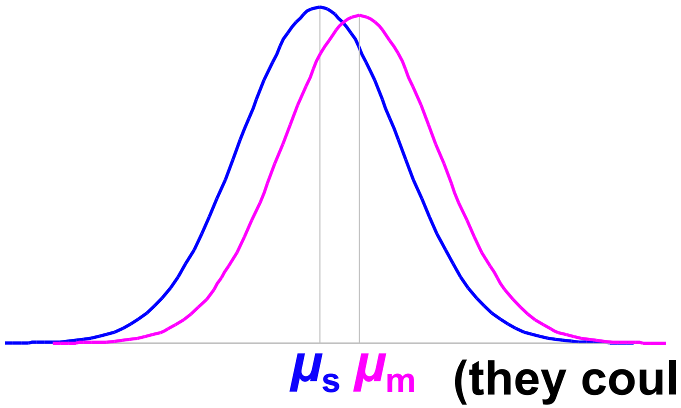
- Goal is to infer population characteristics from sample characteristics



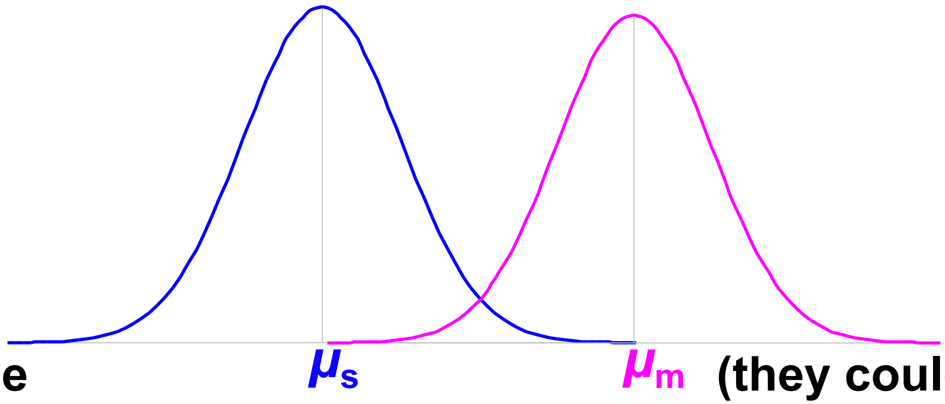
[Howell 2002, p 78]

What Are the Possible Alternatives?

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

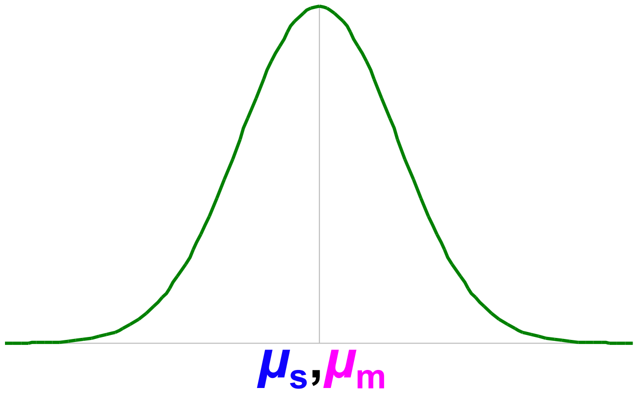


(they could be close together)



(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]

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

$$\left.\begin{array}{l} p \rightarrow q \\ \neg q \\ \hline \rightarrow \neg p \end{array}\right\} \text{modus tollens}$$

[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

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

[Cohen 1994]

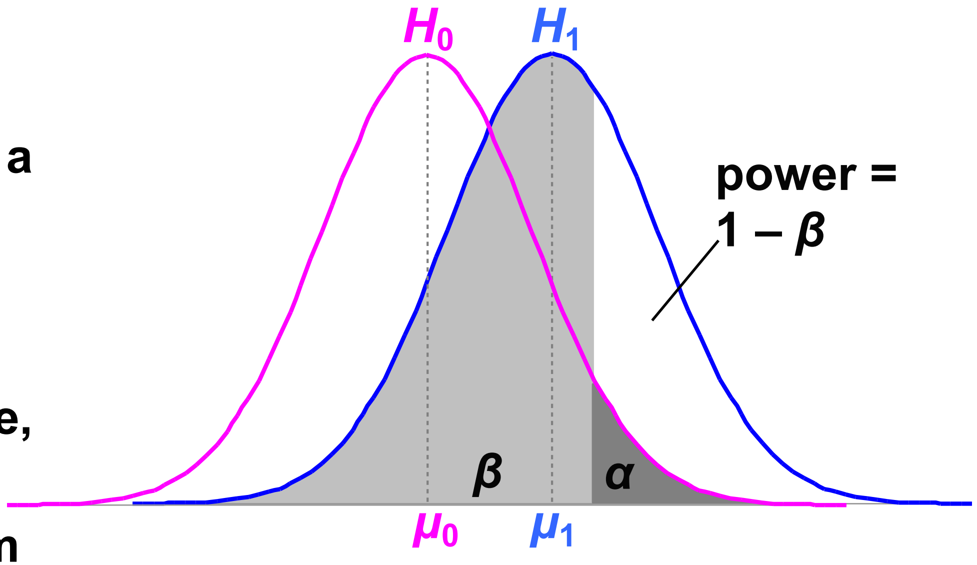
Reproducibility and Inferential Statistics

- 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

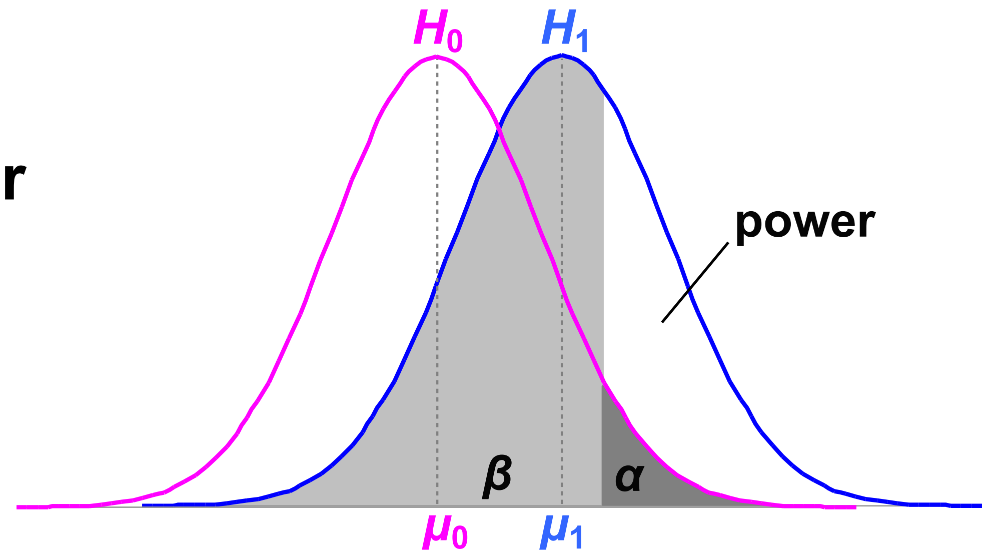
		Decision	
		Reject H_0	Don’t reject H_0
True state of the world	H_0 false	a result! $p = 1 - \beta$ = power	type II error $p = \beta$
	H_0 true	type I error $p = \alpha$	argue H_0? $p = 1 - \alpha$

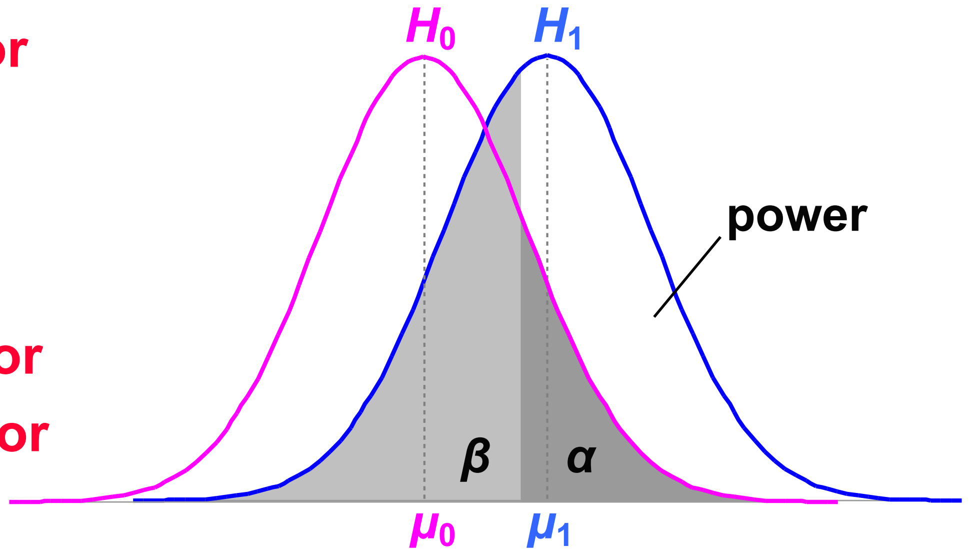
- If H_0 is true:
  - α is probability we make a **type I error**: we think we have a result, but we are wrong
- If H_1 is true:
  - β is probability we make a **type II error**: a result was there, but we missed it
  - **Power** is a more common term than β



Increasing Power by Increasing *α*

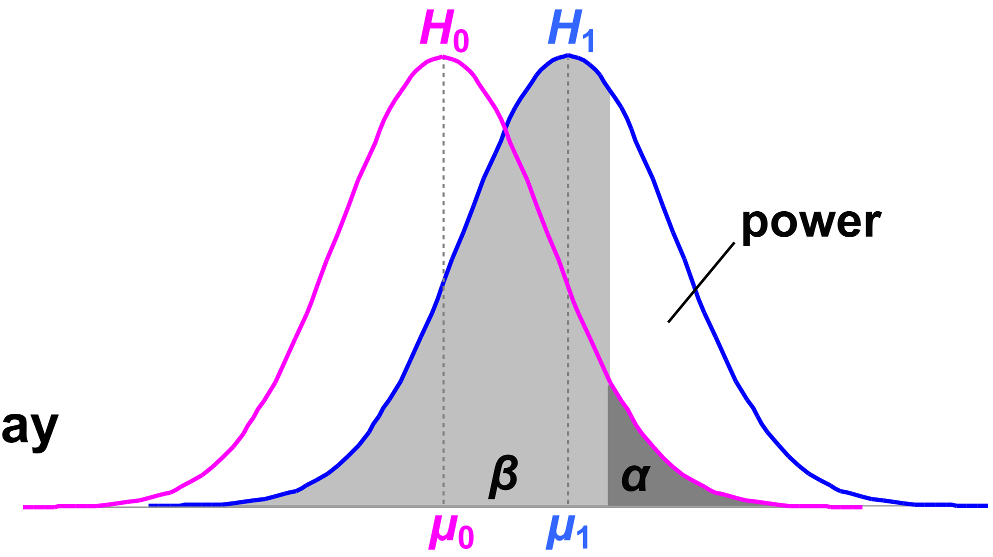
- **Illustrates *α* / power tradeoff**
- **Increasing *α*:**
  - **Increases power**
  - **Decreases type II error**
  - **Increases type I error**
- **Decreasing *α*:**
  - **Decreases power**
  - **Increases type II error**
  - **Decreases type I error**



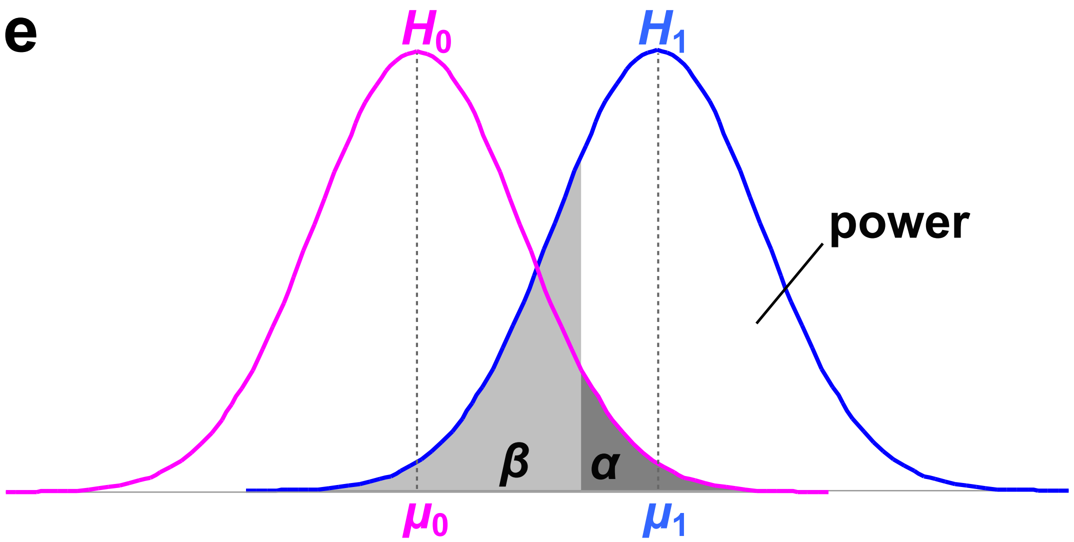


Increasing Power by Measuring a Bigger Effect

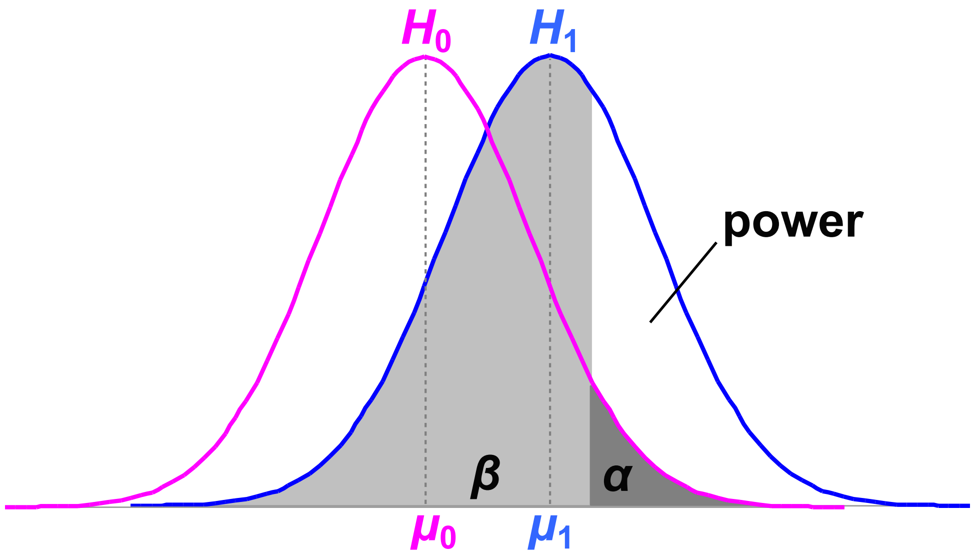
- If the effect size is large:
  - Power increases
  - Type II error decreases
  - α and type I error stay the same



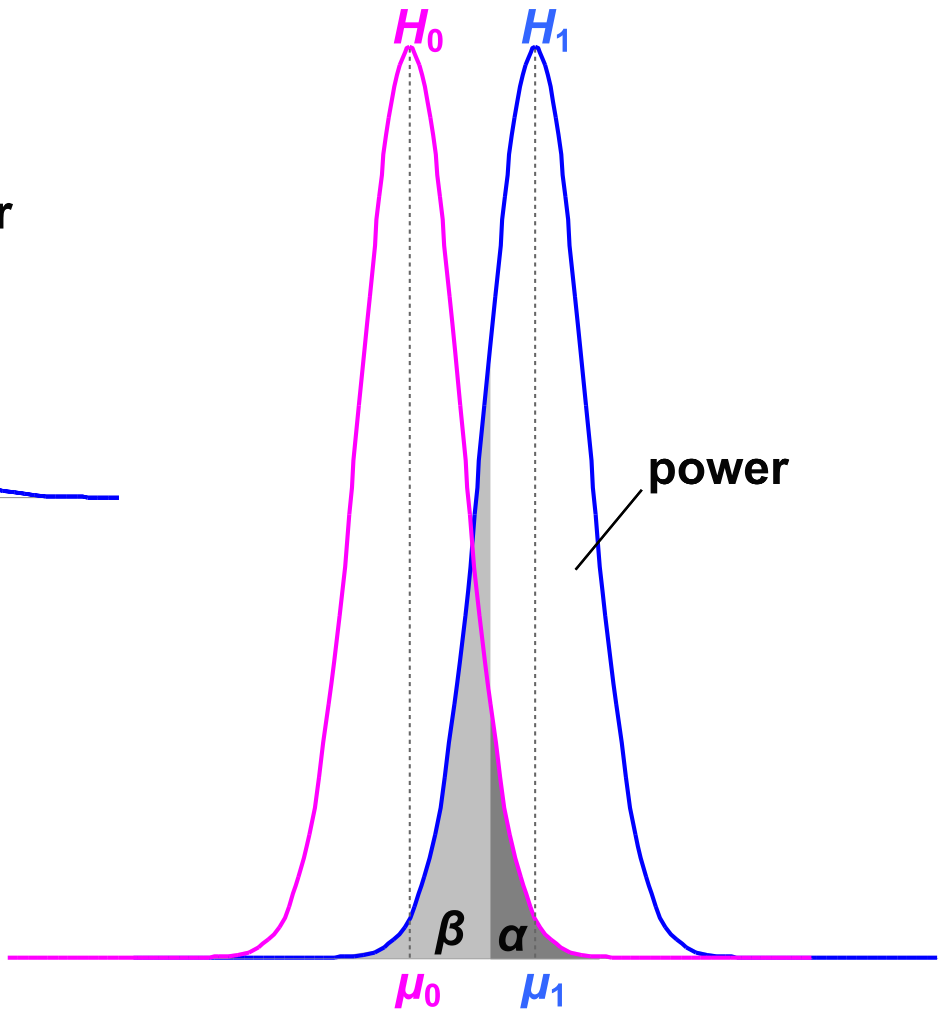
- Unsurprisingly, large effects are easier to detect than small effects



Increasing Power by Collecting More Data

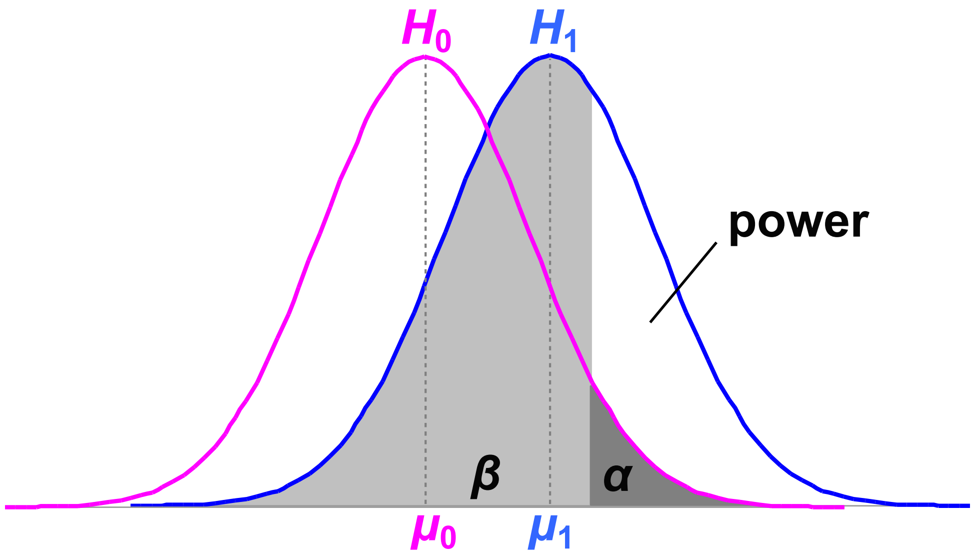


- Increasing sample size (*N*):
  - Decreases variance
  - Increases power
  - Decreases type II error
  - *α* and type I error stay the same
- There are techniques that give the value of *N* required for a certain power level.

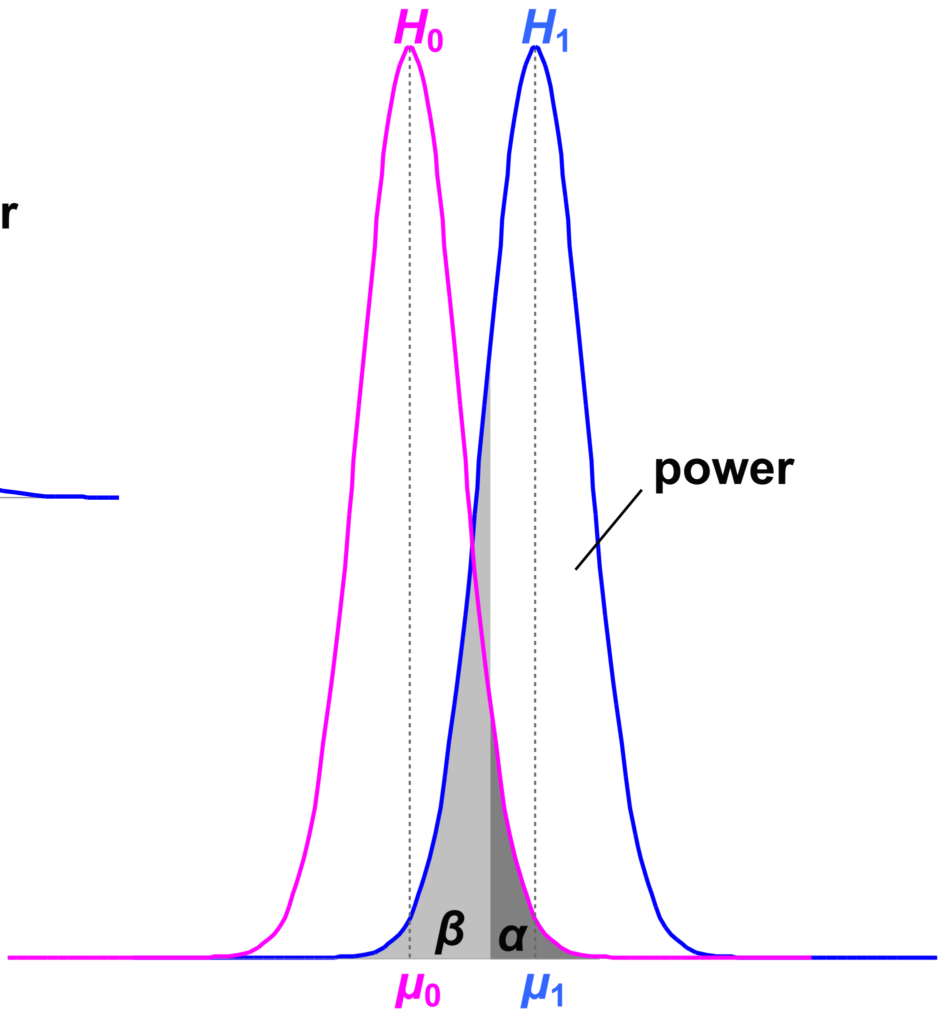


- Here, effect size remains the same, but variance drops by half.

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 relatively new, not many prior studies
  - Can find effect sizes in more mature fields

- Post-hoc power analysis:

$$\text{effect size} = |X_0 - X_1|$$

  - 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:**

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

   - **Number extra samples needed for more powerful result**
   - **Gives “rational basis” for deciding *N***
   - **Cohen [1988] recommends *α* = 0.05, power = 0.80**

2. **Effect size that will be detectable:**

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

3. **Significance level needed:**

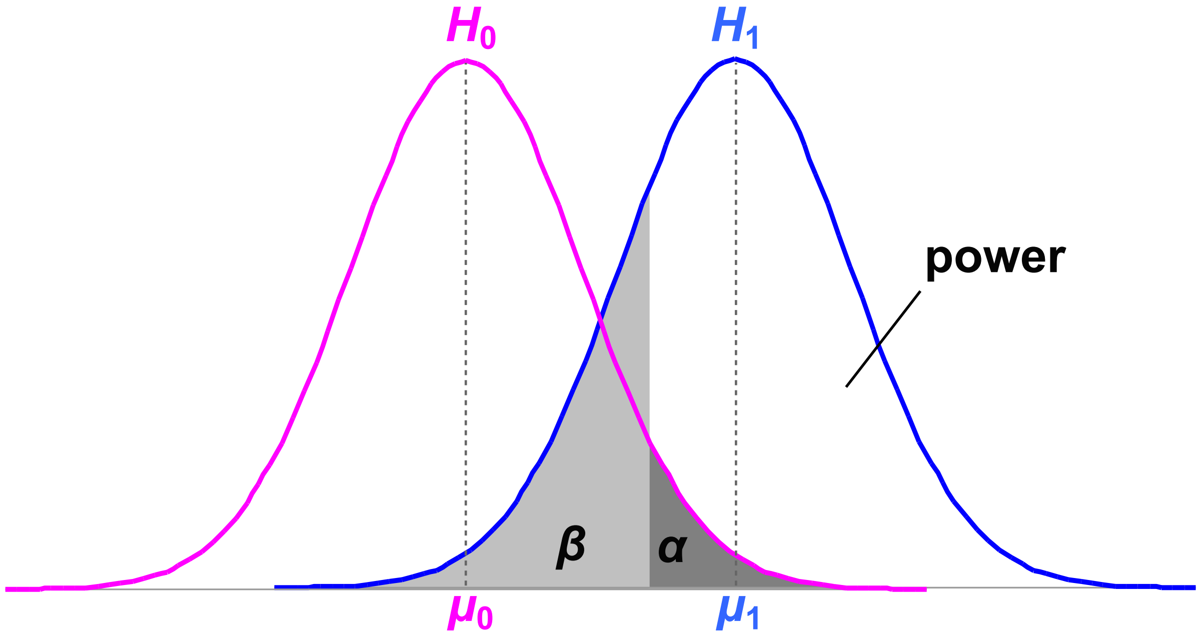
$$\alpha = f(|\mu_0 - \mu_1|, N, \text{power})$$

(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 = 0.80



[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

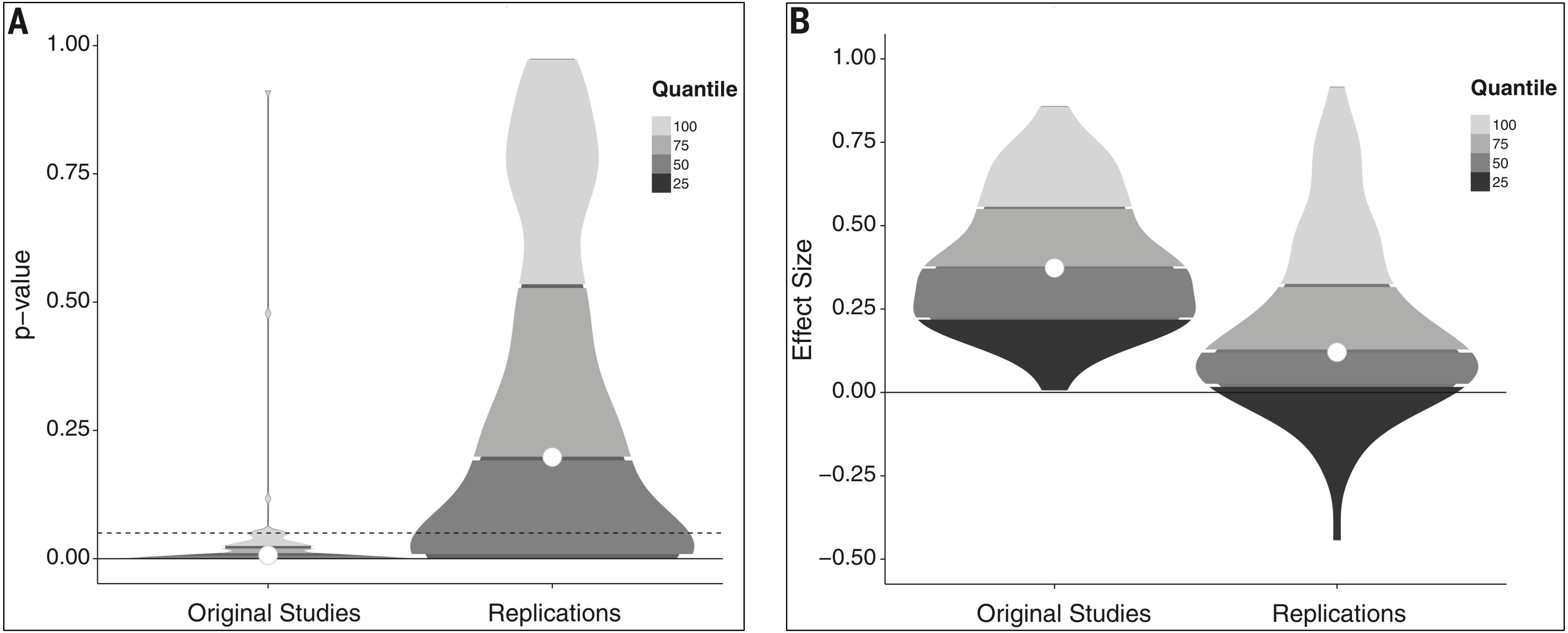
Data collection and processing

- How to measure a replication?
- How to quantify a series of replications?

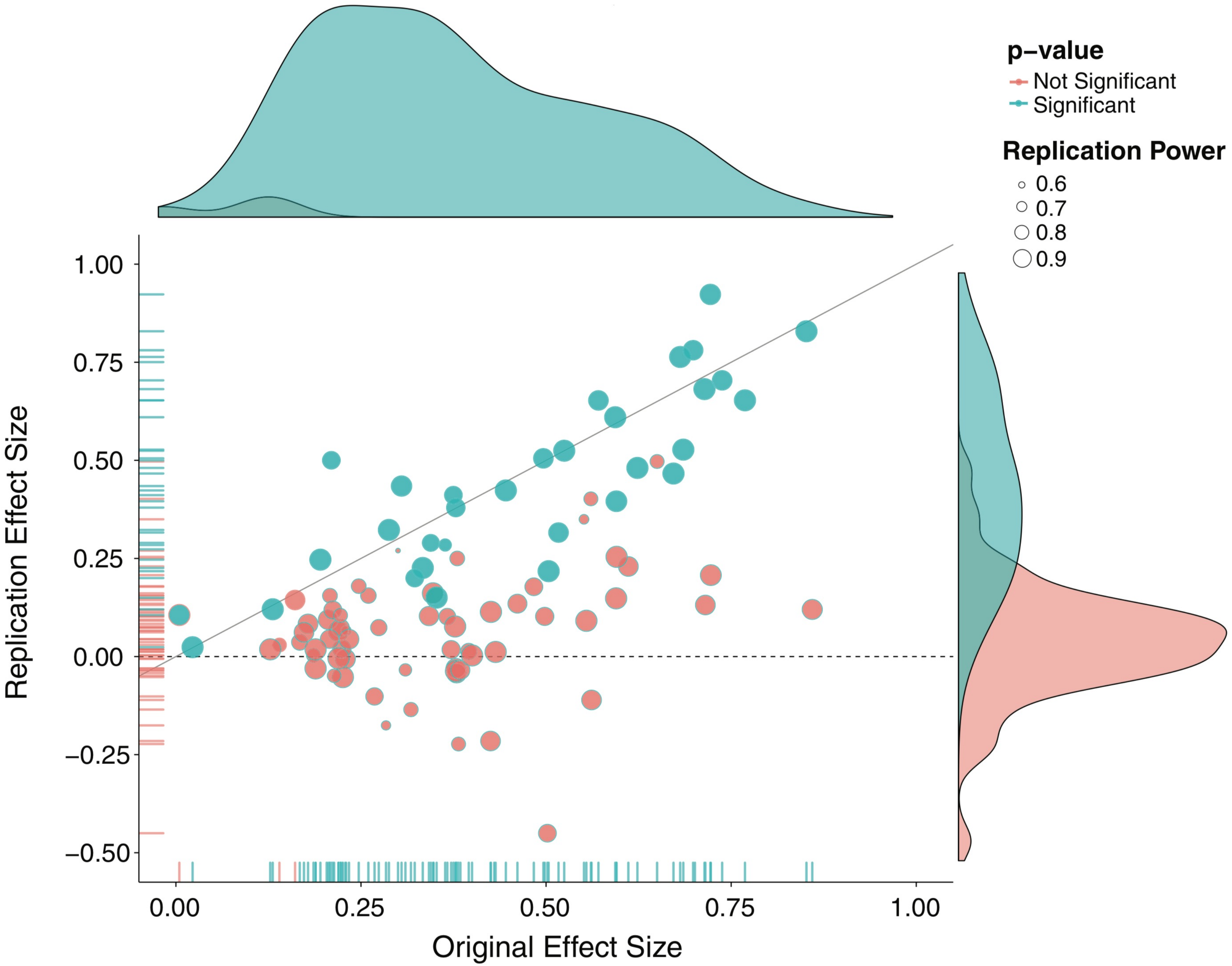
- Each experiment analyzed with standard R packages
- Each analysis performed independently by 2nd team

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

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
Original Effect Size
0.00
0.25
0.50
0.75
1.00

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

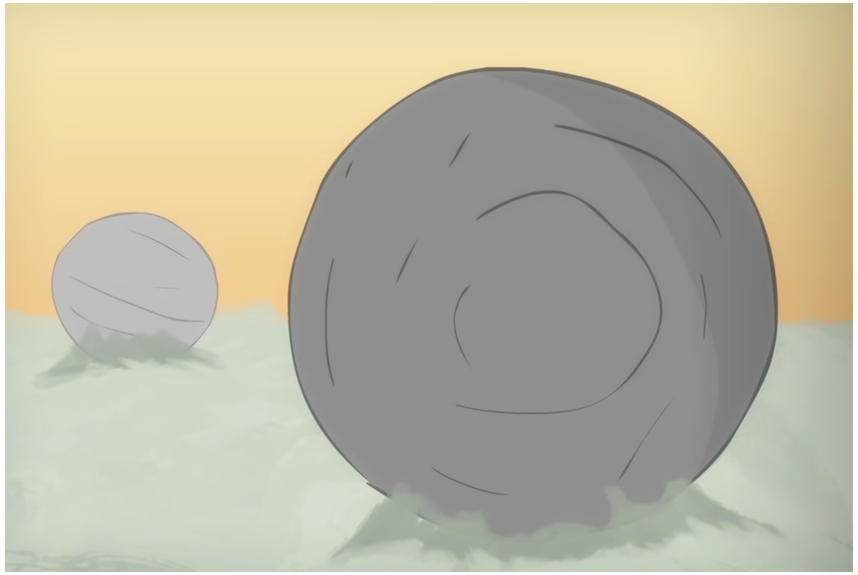
  - Lack of biases likely big part of the reason

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?

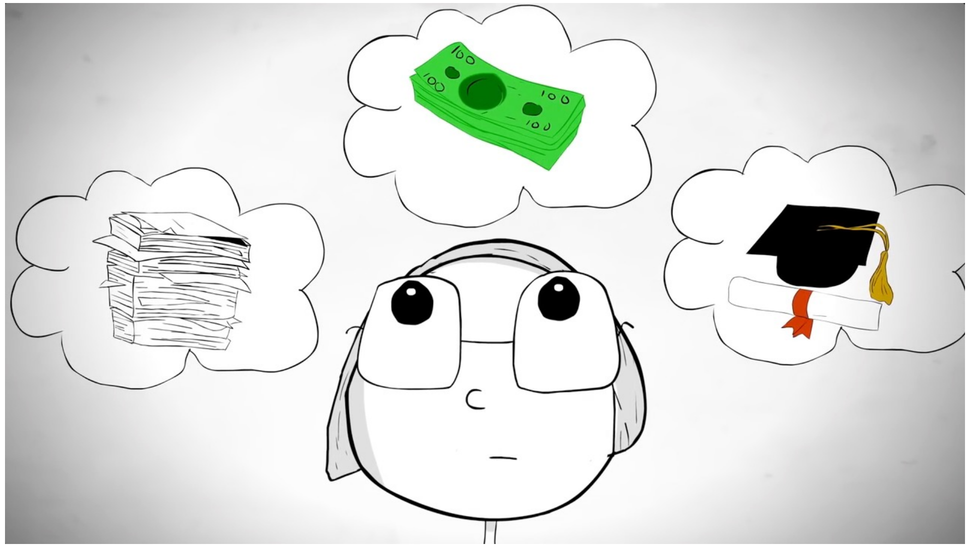
Reasons for Irreproducibility

- 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

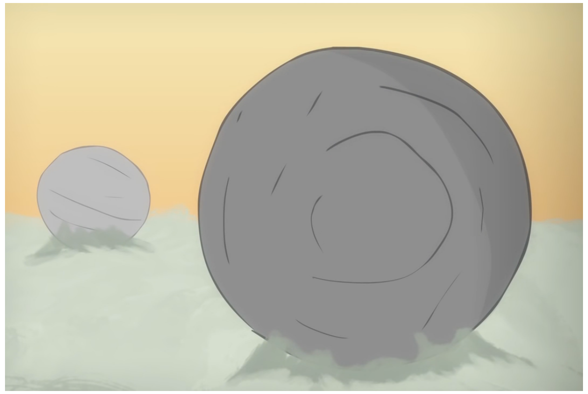
- A study finds A, and the replication study finds A. What does this mean?

  → 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

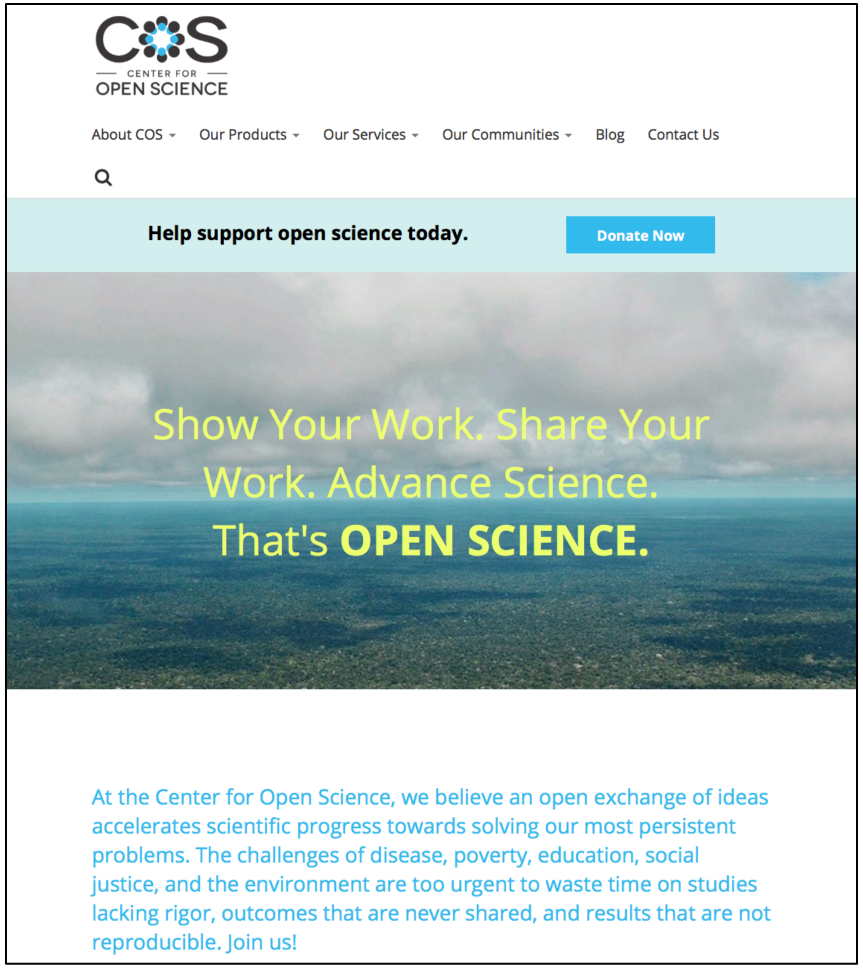
- Describe methods in more detail → easier replication
  - Problem in our field: limited pages
  - Solutions:
    - Additional details in supplementary material, or in associated thesis / dissertation
    - We could adopt longer page limits
    - Main paper in bigger font, methods in smaller font (e.g., *Nature*)
- Upload materials to open repositories → easier replication
  - Data, materials, code
    - Center for Open Science (https://cos.io)
    - 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
  - Channeling 1M USD to pre-registration challenge



Help support open science today.

At the Center for Open Science, we believe an open exchange of ideas accelerates scientific progress towards solving our most persistent problems. The challenges of disease, poverty, education, social justice, and the environment are too urgent to waste time on studies lacking rigor, outcomes that are never shared, and results that are not reproducible. Join us!

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

Research Fellow, Social Science Research Center

Mississippi State University

swan@acm.org

+1-662-325-7507

Mohammed Safayet Arefin

PhD Student, Department of Computer Science and Engineering

Mississippi State University

arefin@acm.org

+1-662-497-6031

Slide Location:

web.cse.msstate.edu/~swan/teaching/tutorials/Swan-VR2021-Tutorial-Replication-Crisis.pdf

Tutorial 2:

The Replication Crisis in Empirical Science: Implications for Human Subject Research in Mixed Reality

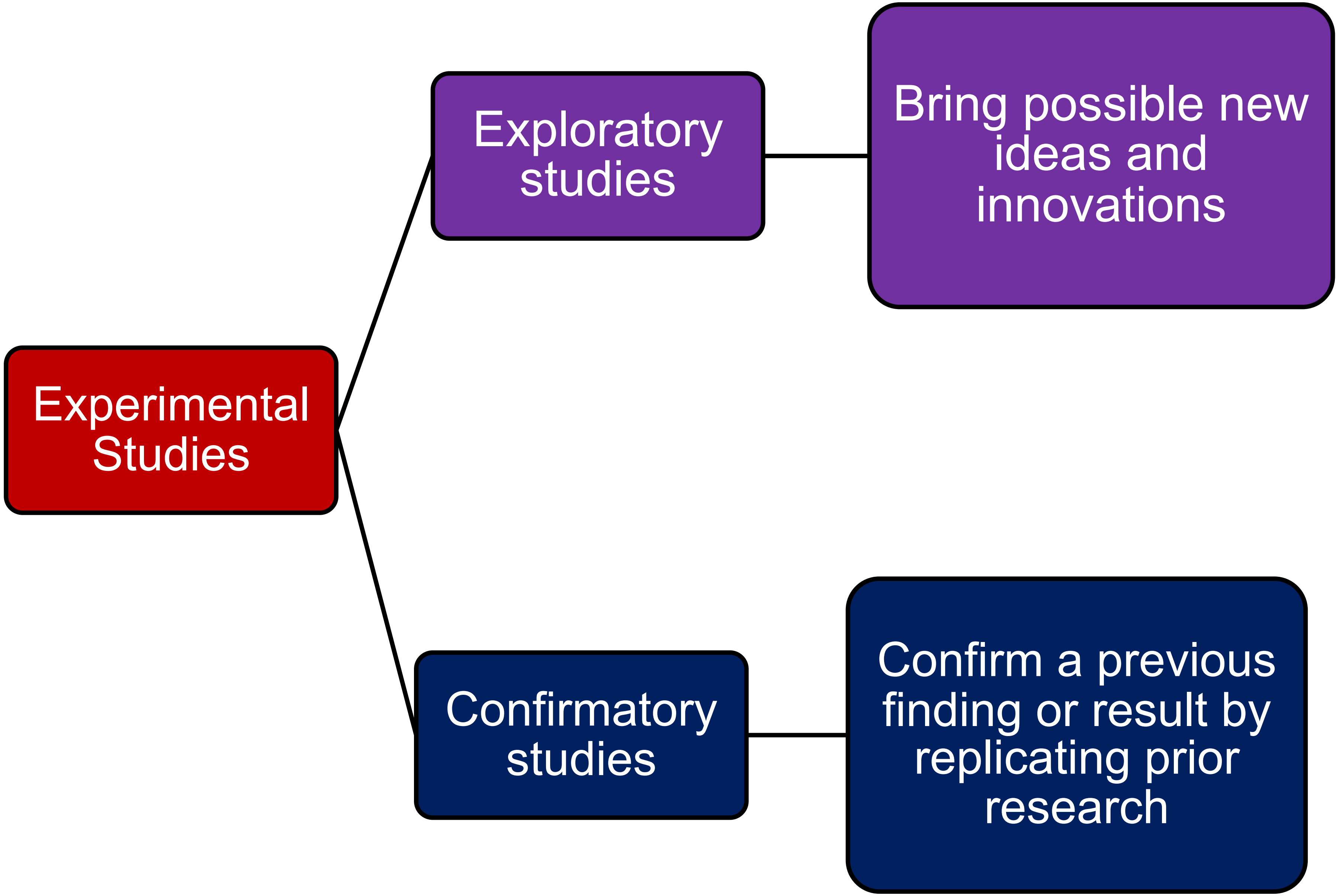
Mohammed Safayet Arefin

Ph.D. Candidate

Computer Science and Engineering

Mississippi State University, USA.

Studies in Empirical Research



Experimental Studies

Exploratory studies

Bring possible new ideas and innovations

Confirmatory studies

Confirm a previous finding or result by replicating prior research

Replication

What is Replication?

According to common understanding,

“Replication is repeating a previous study’s procedure and observing whether the prior finding recurs.”

This definition is **intuitive**, **easy to apply**, and **incomplete**.

❑ Replication is a study for which any outcome would be considered diagnostic evidence about a claim from prior research.

Successful

Successful replication provides evidence of generalizability across the conditions that inevitably differ from the original study

Unsuccessful

Unsuccessful replication indicates that the reliability of the finding may be more constrained than recognized previously.

[Nosek and Errington 2020]

File drawer problem

- A bias toward publishing successful research (i.e., significant results).
- Unsuccessful replications could be stored away in someone’s file drawer!

[Rosenthal 1979]

Replication in Empirical Research

Goals:

- **Verify**
- **Validate**
- **Generalizes**
- **Establishes**

the prior theories, hypotheses, models and findings in the research community.

[Kasper, Søren, Javier, and Jakob 2014]

Replication Typologies

Replication Typologies

Replications Types in Experimental Disciplines

Omar S. Gómez
Universidad Politécnica de Madrid
Boadilla del Monte 28660
Madrid, Spain
ogomez@ieee.org

Natalia Juristo
Universidad Politécnica de Madrid
Boadilla del Monte 28660
Madrid, Spain
natalia@fi.upm.es

Sira Vegas
Universidad Politécnica de Madrid
Boadilla del Monte 28660
Madrid, Spain
svegas@fi.upm.es

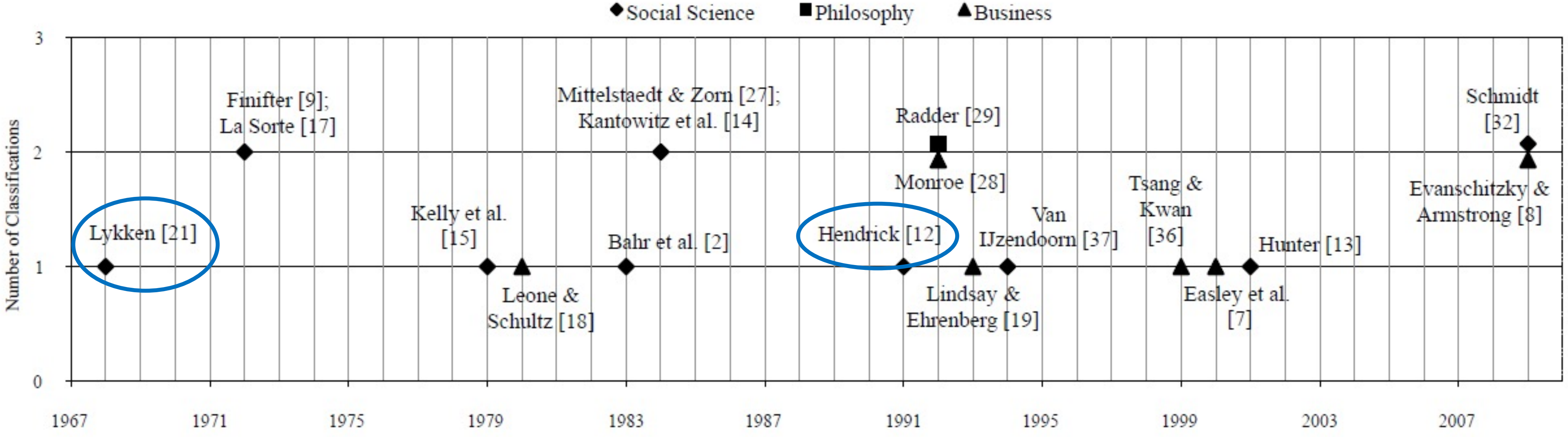
- No general replication classification.
- Replication types rely on the research disciplines.

❑ Selected 18 replication classifications.

❑ Altogether the classifications contain a total of 79 replication types.

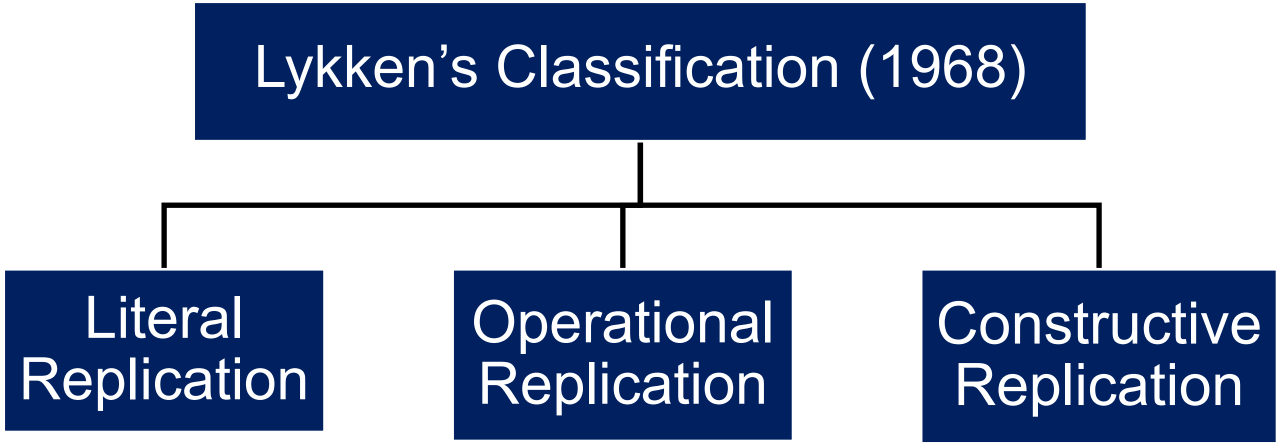
❑ These classifications belong to the fields of Social Science (61%), Business (33%) and Philosophy (6%).

Most often cited classification



Classification Chronology [Gomez, Juristo and Vegas 2010]

Replication Types



Literal Replication:

- It involves exact duplication of the original investigator's sampling procedure, experimental conditions, measuring techniques, and methods of analysis.
- Asking the original investigator to simply run more subjects.

Operational Replication:

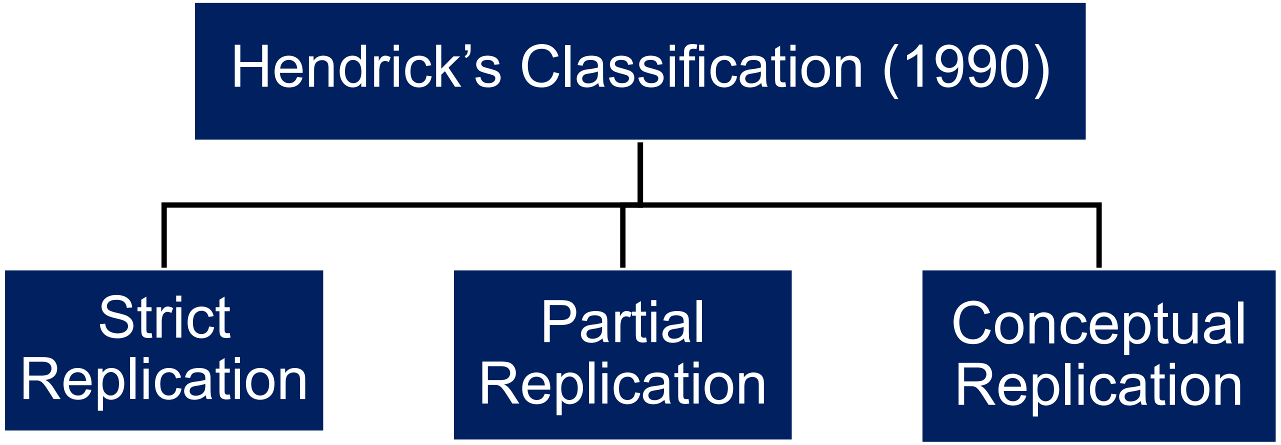
- To test whether the original investigator's "experimental recipe"—will in other hands produce the results that he obtained.

Constructive Replication:

- Replicator formulate his own methods of sampling, measurement, and data analysis to test the empirical "fact" which the original author would claim to have established

[Lykken 1968]

Replication Types



Hendrick's Classification (1990)

Strict Replication

Partial Replication

Conceptual Replication

Strict Replication:

- Aims to replicating the original study as exactly as possible, focusing on the experimental procedure and contextual variables.
- It highly matches the literal replication!

Partial Replication:

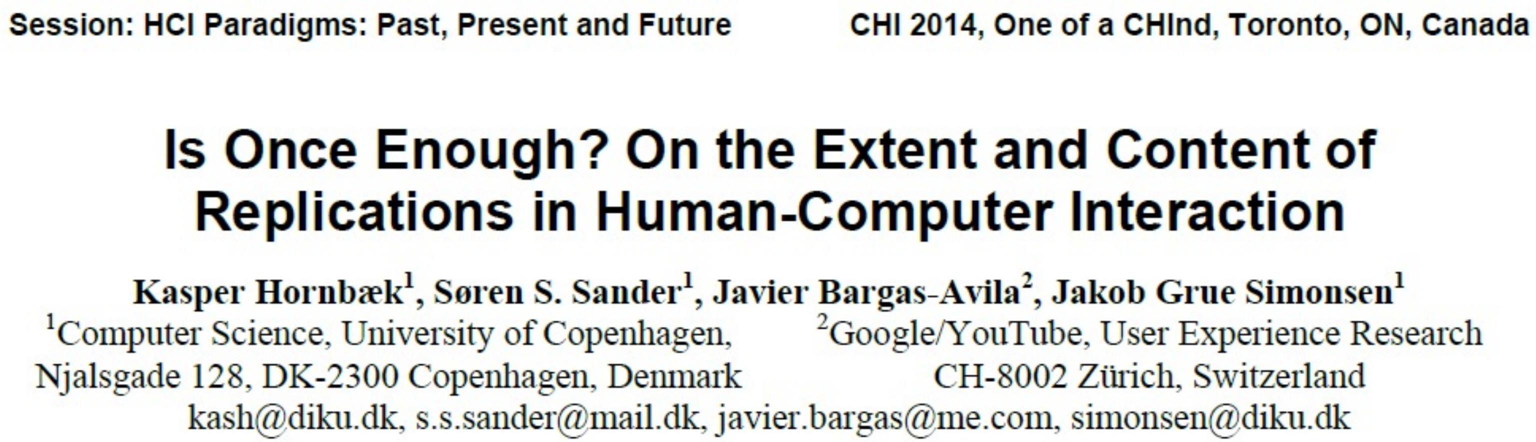
- Some changes (deletion or addition) in part of the experimental variables, while other parts are duplicated as in the original research.
- It is very similar to the operational replication!

Conceptual Replication:

- Use different methods to test the same research question and earlier findings of the prior study.
- It is very similar to the constructive replication!

[Hendrick 1990]

Replication Studies in HCI



Session: HCI Paradigms: Past, Present and Future CHI 2014, One of a CHInd, Toronto, ON, Canada

Is Once Enough? On the Extent and Content of Replications in Human-Computer Interaction

Kasper Hornbæk[1], Søren S. Sander[1], Javier Bargas-Avila[2], Jakob Grue Simonsen[1]

[1]Computer Science, University of Copenhagen,
Njalsgade 128, DK-2300 Copenhagen, Denmark

[2]Google/YouTube, User Experience Research
CH-8002 Zürich, Switzerland

kash@diku.dk, s.s.sander@mail.dk, javier.bargas@me.com, simonsen@diku.dk

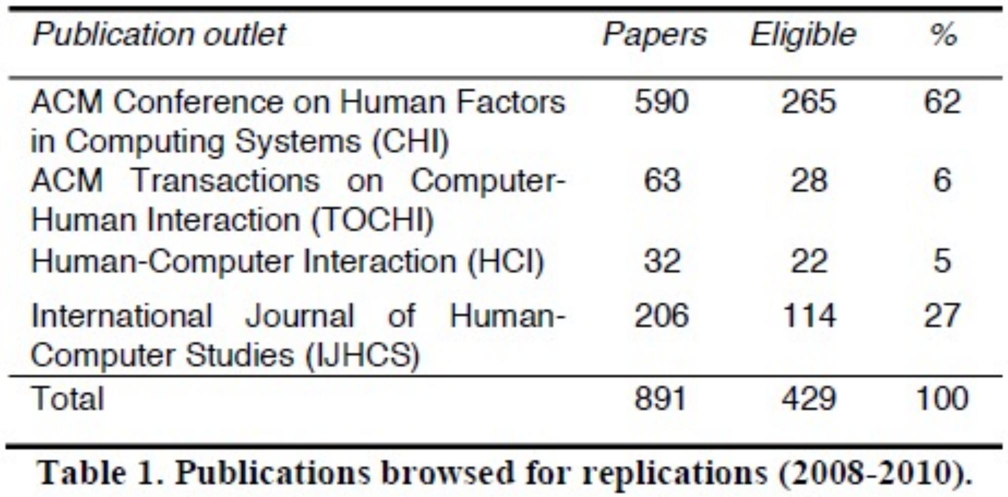
All full papers from the years 2008 to 2010 of a key conference on human-computer interaction and three highly-ranked journals.

Certain criteria for eligibility

- *Be empirical*.
- *Include quantitative data*.
- *Report an experiment*
- *Study human interaction with user interfaces*

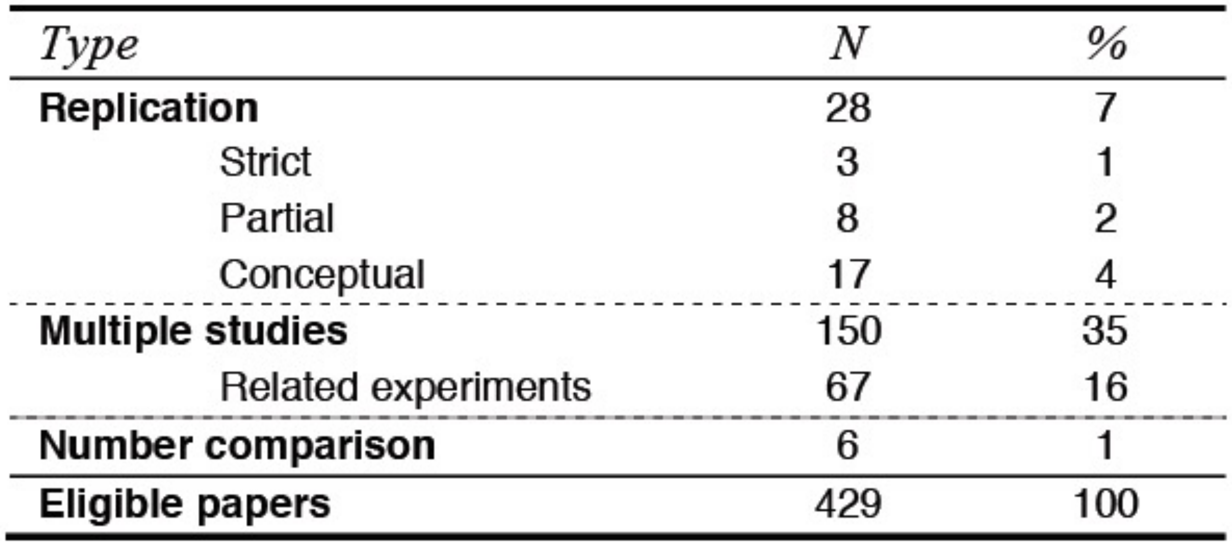
What Counts as a Replication?

An attempt to *confirm, expand, or generalize* an earlier study's findings.

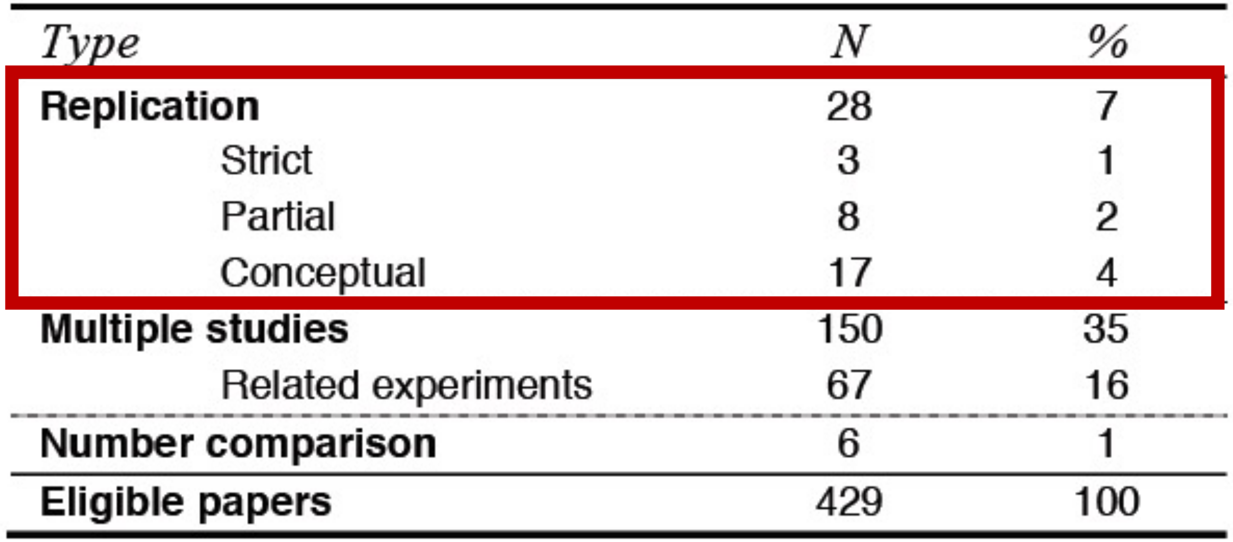


Publication outlet	*Papers*	*Eligible*	*%*
ACM Conference on Human Factors in Computing Systems (CHI)	590	265	62
ACM Transactions on Computer-Human Interaction (TOCHI)	63	28	6
Human-Computer Interaction (HCI)	32	22	5
International Journal of Human-Computer Studies (IJHCS)	206	114	27
Total	891	429	100

Table 1. Publications browsed for replications (2008-2010).



Type	*N*	*%*
Replication	28	7
Strict	3	1
Partial	8	2
Conceptual	17	4
Multiple studies	150	35
Related experiments	67	16
Number comparison	6	1
Eligible papers	429	100



Type	*N*	*%*
Replication	28	7
Strict	3	1
Partial	8	2
Conceptual	17	4
Multiple studies	150	35
Related experiments	67	16
Number comparison	6	1
Eligible papers	429	100

28 replication papers

- 6.5% of the eligible papers
- 3.1% of the full sample

Are earlier findings confirmed?

- 14 papers (50%) performed successful replication
- 3 papers fail to replicate findings
- 4 papers are undefined

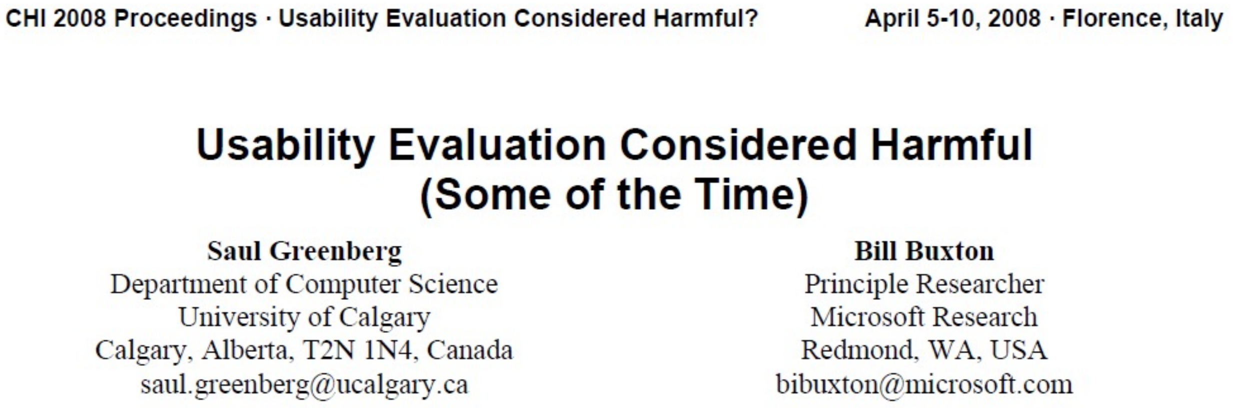
Whose work is replicated?

- 22 studies replicate the work of other researchers
- 2 studies replicate their own work

What do authors' of replications think?

- 13 reported that while their work contained replication.
- Their main goal was not to replicate an original study.
- They emphasized that their main goal was to research new/additional topics and extend the original work.

[Kasper, Søren, Javier, and Jakob 2014]



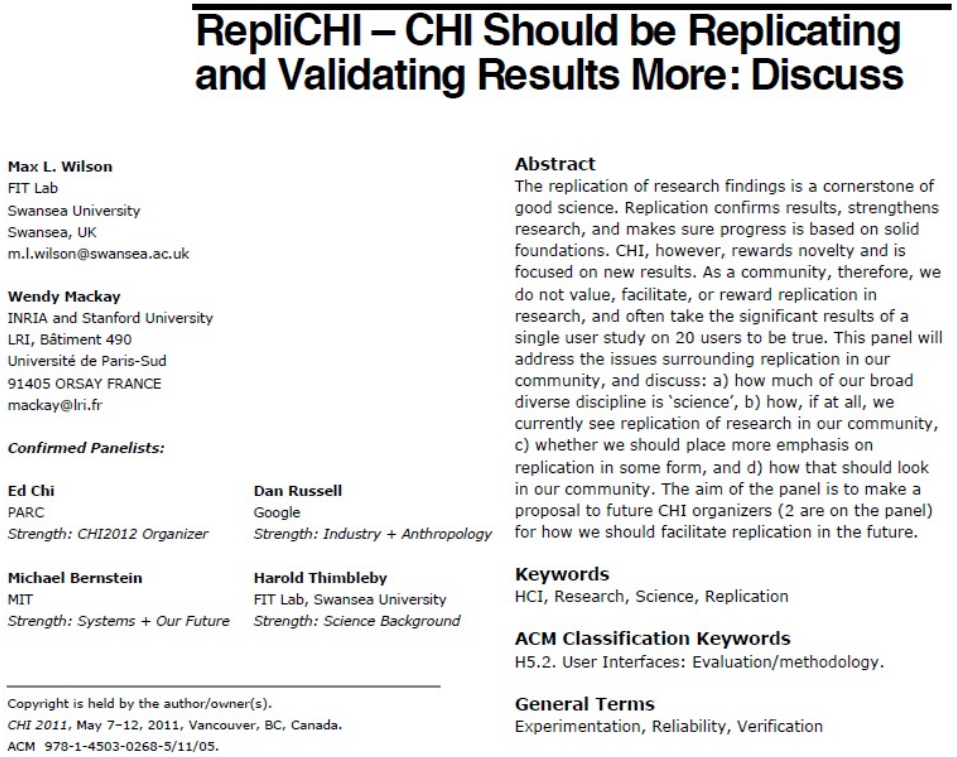
CHI 2008 Proceedings · Usability Evaluation Considered Harmful? April 5-10, 2008 · Florence, Italy

Usability Evaluation Considered Harmful (Some of the Time)

Saul Greenberg
Department of Computer Science
University of Calgary
Calgary, Alberta, T2N 1N4, Canada
saul.greenberg@ucalgary.ca

Bill Buxton
Principle Researcher
Microsoft Research
Redmond, WA, USA
bibuxton@microsoft.com

[Greenberg and Buxton 2008]



RepliCHI – CHI Should be Replicating and Validating Results More: Discuss

Max L. Wilson
FIT Lab
Swansea University
Swansea, UK
m.l.wilson@swansea.ac.uk

Wendy Mackay
INRIA and Stanford University
LRI, Bâtiment 490
Université de Paris-Sud
91405 ORSAY FRANCE
mackay@lri.fr

Confirmed Panelists:

Ed Chi
PARC
Strength: CHI2012 Organizer

Dan Russell
Google
Strength: Industry + Anthropology

Michael Bernstein
MIT
Strength: Systems + Our Future

Harold Thimbleby
FIT Lab, Swansea University
Strength: Science Background

Copyright is held by the author/owner(s).
CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
ACM 978-1-4503-0268-5/11/05.

Abstract

The replication of research findings is a cornerstone of good science. Replication confirms results, strengthens research, and makes sure progress is based on solid foundations. CHI, however, rewards novelty and is focused on new results. As a community, therefore, we do not value, facilitate, or reward replication in research, and often take the significant results of a single user study on 20 users to be true. This panel will address the issues surrounding replication in our community, and discuss: a) how much of our broad diverse discipline is 'science', b) how, if at all, we currently see replication of research in our community, c) whether we should place more emphasis on replication in some form, and d) how that should look in our community. The aim of the panel is to make a proposal to future CHI organizers (2 are on the panel) for how we should facilitate replication in the future.

Keywords

HCI, Research, Science, Replication

ACM Classification Keywords

H5.2. User Interfaces: Evaluation/methodology.

General Terms

Experimentation, Reliability, Verification

[Wilson, Mackay, Chi, Bernstein, Russell, and Thimbleby 2011]

Is replication important for HCI?

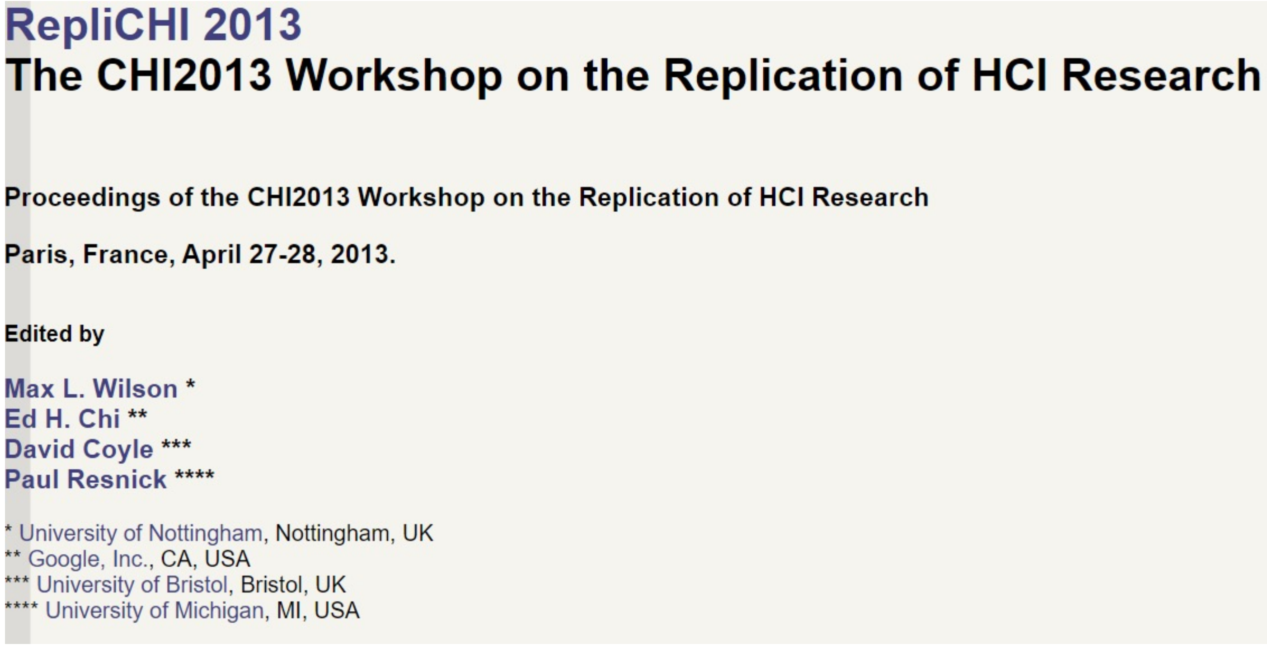
Christian Greiffenhagen
Loughborough University
c.greiffenhagen@lboro.ac.uk

Stuart Reeves
University of Nottingham
stuart@tropic.org.uk

Abstract

Replication is emerging as a key concern within subsections of the HCI community. In this paper, we explore the relevance of science and technology studies (STS), which has addressed replication in various ways. Informed by this literature, we examine HCI's current relationship to replication and provide a set of recommendations and points of clarification that a replication agenda in HCI should concern itself with.

[Christian and Reeves 2013]



RepliCHI 2013
The CHI2013 Workshop on the Replication of HCI Research

Proceedings of the CHI2013 Workshop on the Replication of HCI Research

Paris, France, April 27-28, 2013.

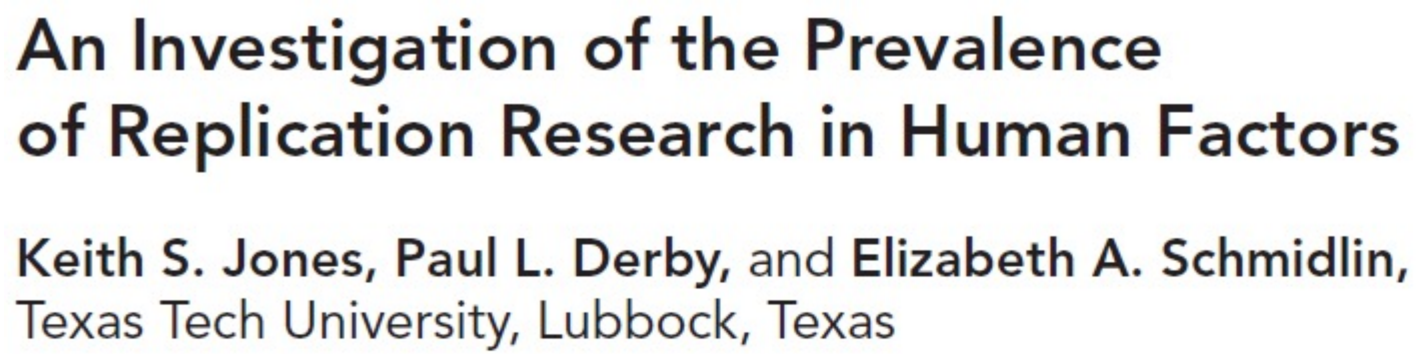
Edited by

Max L. Wilson *
Ed H. Chi **
David Coyle ***
Paul Resnick ****

* University of Nottingham, Nottingham, UK
** Google, Inc., CA, USA
*** University of Bristol, Bristol, UK
**** University of Michigan, MI, USA

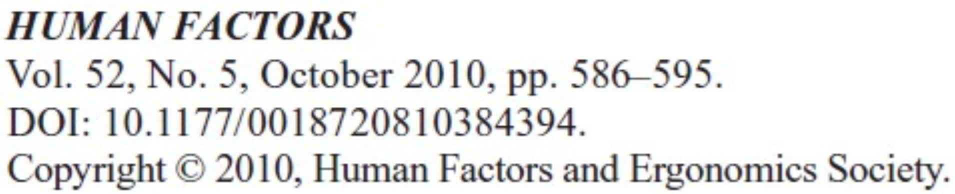
[Link: http://ceur-ws.org/Vol-976/]

Replication Studies in Human Factors Research



An Investigation of the Prevalence of Replication Research in Human Factors

Keith S. Jones, Paul L. Derby, and **Elizabeth A. Schmidlin,**
Texas Tech University, Lubbock, Texas



HUMAN FACTORS
Vol. 52, No. 5, October 2010, pp. 586–595.
DOI: 10.1177/0018720810384394.

Method

Parent articles:

1991 issues of the journal *Human Factors*.

- *8 articles were selected*

Child articles:

- articles that cited the parent articles between 1991 and September 2006.

❑ Compared each child article against parent article to determine whether the child article replicated its parent article. Nonempirical child articles were omitted.

Were human factors studies replicated? **Yes!**

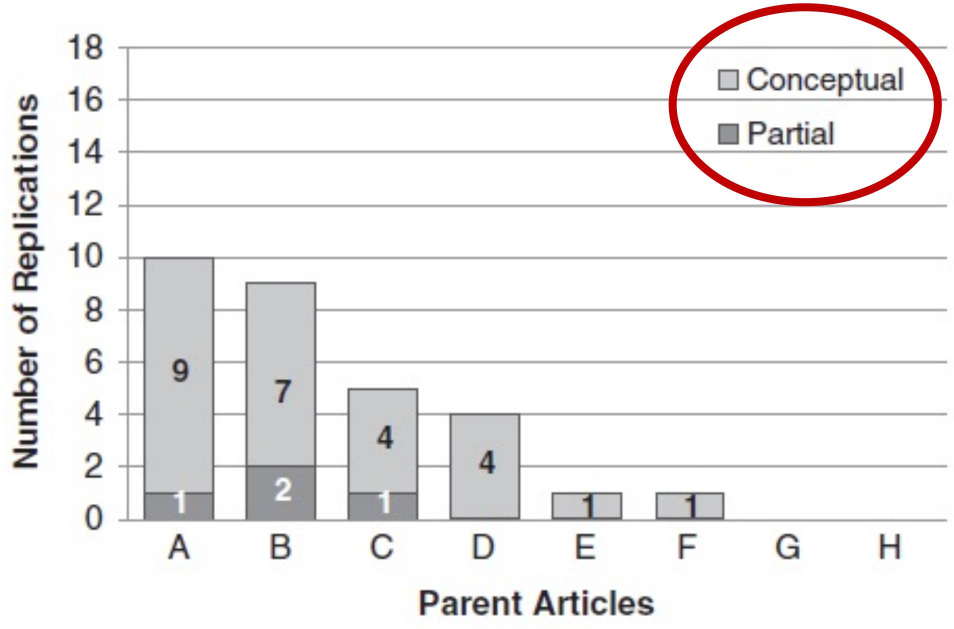
Number of replications: **30**

How frequently were studies replicated?

0 to 10 times (*M* = 3.75, *SE* = 1.41).

Were the replications successful?

- 28 (93%) were successful replications
- 2 unsuccessful replications

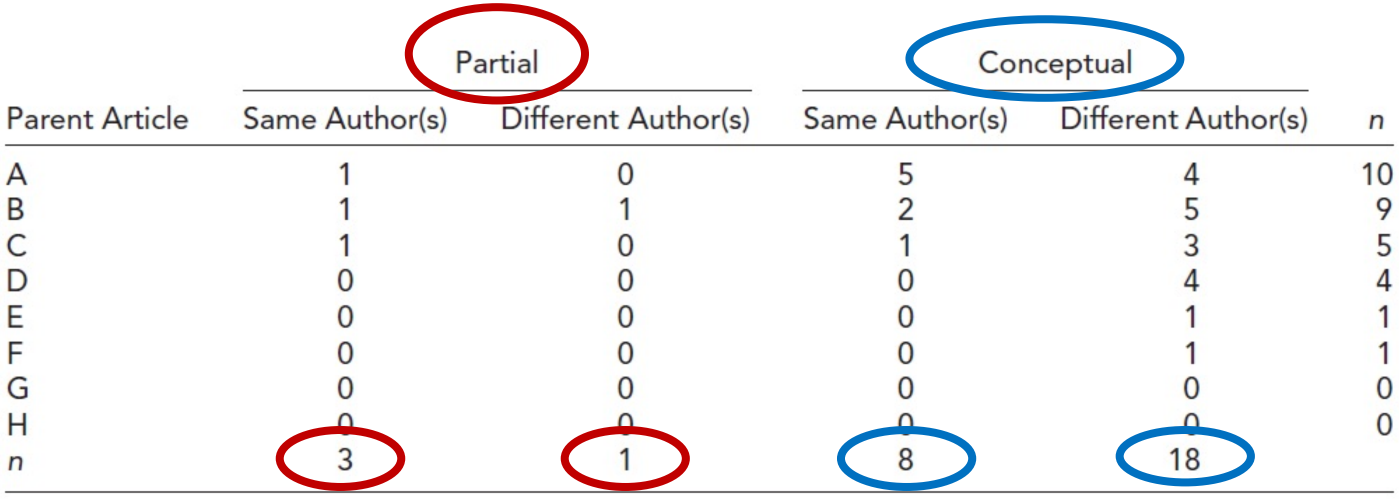


Any variation of the word replication (e.g., replicate, replicated)?

- 6 of the 30 replications (20%) included the words *replication* or *replicated* within the text of the article.
- 24 replications (80%) did not include any language about replication.
- Authors may not have known that their research was a replication or may not have considered their research to be a replication.

Who conducted the replications?

- 11 of the 30 replications (37%) were conducted by original author(s) or coauthor(s).
- Other authors conducted 19 of the 30 total replications (63%).



	Partial		Conceptual		
Parent Article	Same Author(s)	Different Author(s)	Same Author(s)	Different Author(s)	*n*
A	1	0	5	4	10
B	1	1	2	5	9
C	1	0	1	3	5
D	0	0	0	4	4
E	0	0	0	1	1
F	0	0	0	1	1
G	0	0	0	0	0
H	0	0	0	0	0
n	3	1	8	18	

Key Points:

- Replications exist within the human factor's literature.
- Exact replications are rare, whereas conceptual and partial replications are more prevalent.
- Replications are not always labeled as replications, so it can be difficult to identify them.

[Jones, Derby, & Schmidlin 2010]

Replication Studies in Augmented and Virtual Reality

To our knowledge,

No research paper has systematically investigated replications in AR/VR research domain.

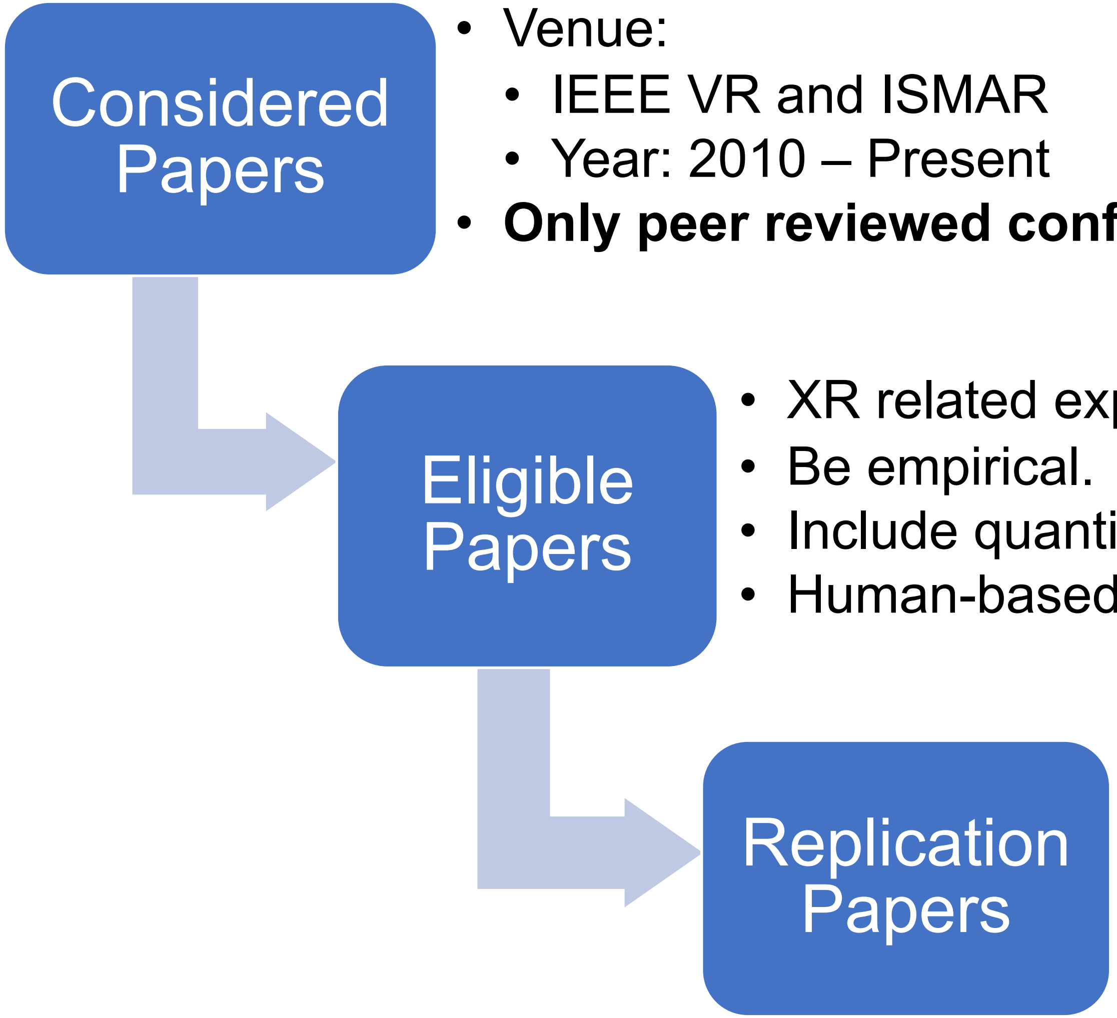
Therefore, we do not know :

(a) the extent of replications in AR/VR

(b) the content of those replications.

Replication Studies in Augmented and Virtual Reality

Our Replication Survey (Initial Approach)



Considered Papers

- Venue:
  - IEEE VR and ISMAR
  - Year: 2010 – Present
- **Only peer reviewed conference papers are considered**

Eligible Papers

- XR related experiment
- Be empirical.
- Include quantitative data.
- Human-based study

Mixed Reality User-Based Studies

Replication Papers

- Must reference the original study.
- An attempt to *confirm, expand, or generalize* an earlier study's findings.

Replication Studies in Augmented and Virtual Reality

Replication Survey (Initial Approach)

Search Process: Eligible Papers to Replication Papers

- **Searching with keywords:**
  - Searching with ***"replication"*** keyword.
    - It is difficult to find that a paper contains "replication" word which is working on replication.
  - We can use other keywords:" reproduce", "follow up study", "earlier experiment", "similar to", "previous experiment", "reproduce results", "inspired by", "adapted from".
  - Check the paper's **"abstract", "Introduction", "Experimental Task" and "Discussion**" section.

- **Further Exploring the Eligible Papers:**
  - The replication paper must reference the original study.
  - Must collect data related to the original study.
  - Not necessary to have replication keyword
  - The replication paper must express intent to do at least one of them:
    - Confirms / expand / generalizes the finding of original study
  - Only comparing results with original work does not count as replication. Needs to prove **experimentally**.

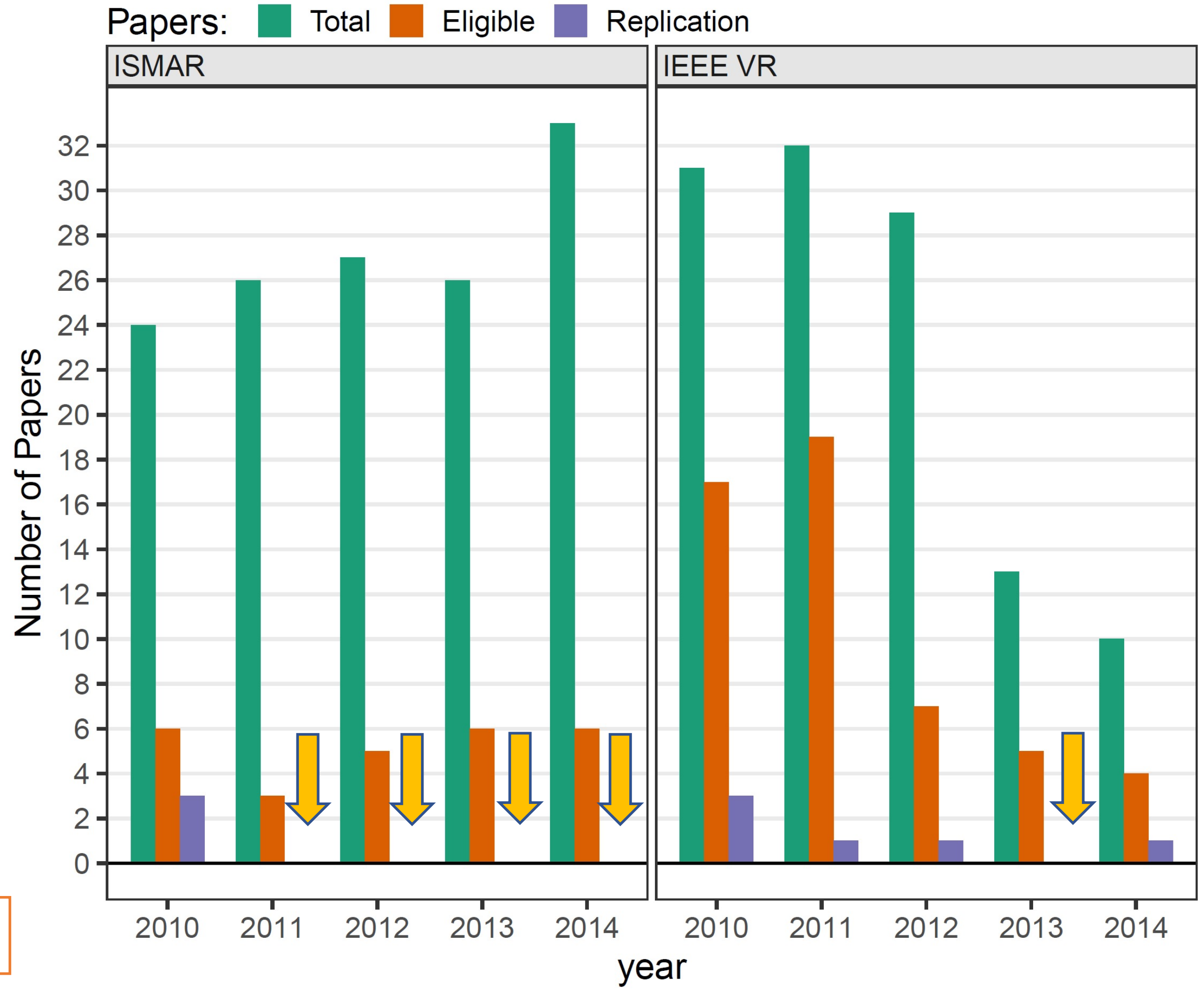
Replication Survey

Initial results

Were XR studies replicated? **Yes!**

Publication Venue	Years	Total Papers	Eligible Papers	Replication Papers
ISMAR	2010-2014	136	26 **(19.11% of total paper)**	3 **(2.2% of total paper)**
IEEE VR	2010-2014	115	52 **(45.22% of total paper)**	6 **(5.2% of total paper)**
		251	78 **(31.07% of total paper)**	**9** **(3.59% of total paper)**

HCI: 3.1% of the full sample



- 5 years (2010-2014).
- Only peer reviewed conference is considered (**No journal publication**).

Future:

We will extend our survey with more publication's venues and years.

Experimental Environment	Replication Papers
AR	4
VR	5

Replication Type	Replication Papers
Strict	0 (0%)
Partial	3 (33.33%)
Conceptual	6 (66.67%)

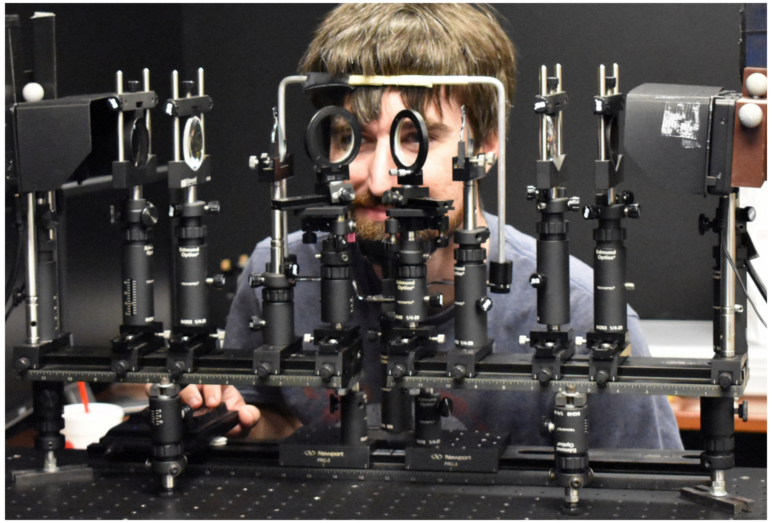
Replication Work

- Joseph L. Gabbard, Divya Gupta Mehra, and J. Edward Swan II. **Effects of AR Display Context Switching and Focal Distance Switching on Human Performance**. *IEEE Transactions on Visualization and Computer Graphics*, 25(6):2228–2241, May 2018. DOI: **10.1109/TVCG.2018.2832633**.

Partial Replication

Same Experimental
- Task
- Procedure

Deliberate Modification:
- Extend experimental variables
- Different AR displays
- Complete experimental design



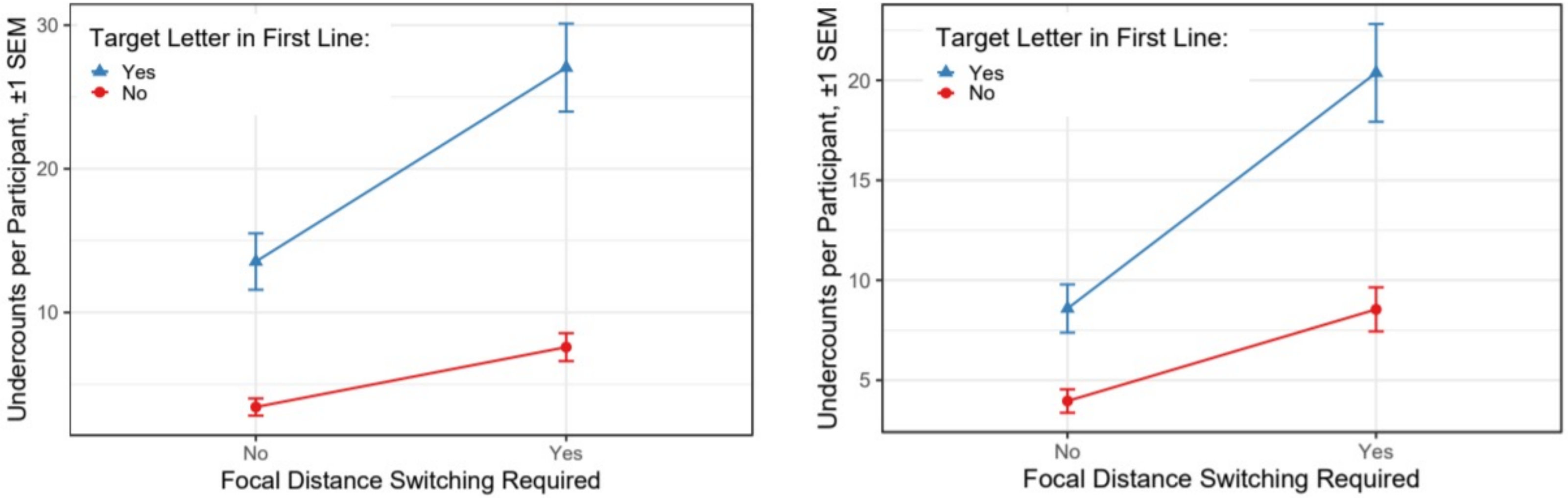
- Mohammed Safayet Arefin, Nate Phillips, Alexander Plopski, Joseph L. Gabbard, and J. Edward Swan II. **Impact of AR Display Context Switching and Focal Distance Switching on Human Performance: Replication on an AR Haploscope**. In *Abstracts and Workshops Proceedings, IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR 2020)*, IEEE Computer Society, March 2020. DOI: 10.1109/VRW50115.2020.00137

Replication Results

- Successfully replicate experiment of Gabbard, Mehra and Swan II (2018).

❑ Context switching and focal distance switching have **significant impact on user performance and eye-fatigue**

❑ Context switching and focal distance switching are **general optical see-through AR user interface design issues.**



(a) Data from Gabbard et al. [2]; 24 participants.

(b) Data collected from AR haploscope (Figure 1); 24 participants.

❑ The results are consistent with the hypothesis that these findings broadly generalize to optical see-through AR user interfaces.

References

- Nosek, B. A., & Errington, T. M. (2020). What is replication?. *PLoS biology*, *18*(3). https://doi.org/10.1371/journal.pbio.3000691
- Omar S. Gómez, Natalia Juristo, and Sira Vegas. 2010. Replications types in experimental disciplines. In Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '10). Association for Computing Machinery, New York, NY, USA, Article 3, 1–10. DOI: https://doi.org/10.1145/1852786.1852790
- Hendrick, C. (1990). Replications, strict replications, and conceptual replications: Are they important? *Journal of Social Behavior & Personality, 5*(4), 41–49.
- Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*(3, Pt.1), 151–159. https://doi.org/10.1037/h0026141
- Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. 2014. Is once enough? on the extent and content of replications in human-computer interaction. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14). Association for Computing Machinery, New York, NY, USA, 3523–3532. DOI: https://doi.org/10.1145/2556288.2557004
- Max L. Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, Dan Russell, and Harold Thimbleby. 2011. RepliCHI - CHI should be replicating and validating results more: discuss. In CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11). Association for Computing Machinery, New York, NY, USA, 463–466. DOI: https://doi.org/10.1145/1979742.1979491
- Christian Greiffenhagen, Stuart Reeves. 2013. Is replication important for HCI? In workshop on replication in HCI (RepliCHI), SIGCHI Conference on Human Factors in Computing Systems (CHI), Paris, France.
- Saul Greenberg and Bill Buxton. 2008. Usability evaluation considered harmful (some of the time). In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08). Association for Computing Machinery, New York, NY, USA, 111–120. DOI: https://doi.org/10.1145/1357054.1357074
- Jones, K. S., Derby, P. L., & Schmidlin, E. A. (2010). An investigation of the prevalence of replication research in human factors. Human factors, 52(5), 586–595. https://doi.org/10.1177/0018720810384394.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86, 638–641.