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

J. Edward Swan II
Mississippi State University
Monday, 4 October 2021
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)
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

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

[Hen Thom 2017]
Evidence

• Replication studies conducted in biomedicine, psychology

• Survey data, based on question:
  • “Have you failed to reproduce somebody else’s experiment?”

<table>
<thead>
<tr>
<th>Field</th>
<th>% Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry</td>
<td>87%</td>
</tr>
<tr>
<td>Biology</td>
<td>77%</td>
</tr>
<tr>
<td>Physics / Engineering</td>
<td>69%</td>
</tr>
<tr>
<td>Medicine</td>
<td>67%</td>
</tr>
<tr>
<td>Earth / Environment</td>
<td>64%</td>
</tr>
<tr>
<td>Other</td>
<td>62%</td>
</tr>
</tbody>
</table>

[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

<table>
<thead>
<tr>
<th>Sponsor</th>
<th>%Replicated</th>
<th>Number Replicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayer</td>
<td>21%</td>
<td>14/67</td>
</tr>
<tr>
<td>Amgen</td>
<td>11%</td>
<td>6/53</td>
</tr>
<tr>
<td>National Institute for Neurological Disorders and Stroke</td>
<td>8%</td>
<td>1/12</td>
</tr>
<tr>
<td>ALS Therapy Development Institute</td>
<td>0%</td>
<td>0/47</td>
</tr>
<tr>
<td>Reproducibility Project: Psychology</td>
<td>36%</td>
<td>35/97</td>
</tr>
</tbody>
</table>
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 ($1 - \beta \text{ (power)} \geq 0.80$)
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

<table>
<thead>
<tr>
<th>Original Study Result Characteristics</th>
<th>Replication Study Result Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$ value</td>
<td>$p$ value</td>
</tr>
<tr>
<td>effect size</td>
<td>effect size</td>
</tr>
<tr>
<td>$df$ or sample size</td>
<td>$df$ or sample size</td>
</tr>
<tr>
<td>result importance rating</td>
<td>power</td>
</tr>
<tr>
<td>result surprisingness rating</td>
<td>replication challenge rating</td>
</tr>
<tr>
<td>experience, expertise rating of original team</td>
<td>experience, expertise rating of replicating team</td>
</tr>
<tr>
<td></td>
<td>replication quality rating</td>
</tr>
</tbody>
</table>
Results
Results
Results by %Replicated ($p \leq 0.05$)

- Initial strength of evidence predicts replication success

<table>
<thead>
<tr>
<th>Original Strength of Evidence</th>
<th>%Replicated ($p \leq 0.05$)</th>
<th>Number Replicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p \leq 0.001$</td>
<td>63%</td>
<td>20/32</td>
</tr>
<tr>
<td>$p \leq 0.02$</td>
<td>41%</td>
<td>26/63</td>
</tr>
<tr>
<td>$0.02 \leq p \leq 0.04$</td>
<td>26%</td>
<td>6/23</td>
</tr>
<tr>
<td>$0.04 \leq p$</td>
<td>18%</td>
<td>2/11</td>
</tr>
</tbody>
</table>

- Cognitive psychology more successful than social psychology

<table>
<thead>
<tr>
<th>Sub-Discipline</th>
<th>%Replicated ($p \leq 0.05$)</th>
<th>Number Replicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Psychology</td>
<td>50%</td>
<td>21/42</td>
</tr>
<tr>
<td>Social Psychology</td>
<td>25%</td>
<td>14/55</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Effect Type</th>
<th>%Replicated ($p \leq 0.05$)</th>
<th>Number Replicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Effect</td>
<td>47%</td>
<td>23/49</td>
</tr>
<tr>
<td>Interaction Effect</td>
<td>22%</td>
<td>8/37</td>
</tr>
</tbody>
</table>
**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 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 $\alpha$ 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) → 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](https://cos.io))
    - 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


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


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

Slide Location:
web.cse.msstate.edu/~swan/teaching/tutorials/Swan-ISMAR2021-WoRXR-Workshop-Replication-Crisis.pdf
The Replication Crisis, Reproducibility, and the Reproducibility Project in Psychology

J. Edward Swan II
Mississippi State University
Monday, 4 October 2021