Why Do So Many Studies Fail to Replicate?

Gray Matter
By JAY VAN BAEL
MAY 27, 2016
Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition

“I SEE a train wreck looming,” warned Daniel Kahneman, an eminent psychologist, in an
More than half of psychology papers are not reproducible

Initiative to replicate findings of 100 prominent studies raises further questions about health of discipline

August 27, 2015

By Paul Jump Twitter: @PaulJump
How to replicate
### What is a replication?

<table>
<thead>
<tr>
<th>Duplication</th>
<th>Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verify research results</td>
<td>Test the robustness of the original research results</td>
</tr>
<tr>
<td>exact same data set</td>
<td>new data</td>
</tr>
<tr>
<td>exact same methods</td>
<td>new models</td>
</tr>
</tbody>
</table>
Which study should I pick?

**Relevant research with impact**

**Results widely accepted but never checked**

**Outdated measures**

**Missing control variables**

---

The perfect replication project

Abstract
I’m the perfect replication project because I combine all these, or at least most of these, features: interesting & relevant questions, results that are accepted but have never been checked, fail to control for important variables, use out-dated measurements, make you wonder if the results apply in different contexts, I’m pointed at in “limitations” and “future research” sections of articles, I’m in an area ‘ripe for replication’. 

Keywords: replication, relevant, improvement
Examples of a ‘good pick’


**Argument**: high debt is associated with lower growth

**Impact**:  
- high journal (The American Economic Review)  
- research was used by governments to justify austerity measures
Practical steps in a replication study

1. Select paper
2. Access data & code
3. Identify each variable
4. Reproduce tables, figures

If you got to this point, you completed a duplication.
Practical steps in a replication study (II)

6 Add value
   • new data
   • new variables
   • new model specifications
   • theoretical contributions

7 Compare

8 Get feedback from peers

9 Journal submission

You now completed a full replication!
Communicating failed replications
“We ... find that coding errors, selective exclusion of available data, and unconventional weighting of summary statistics lead to serious errors” (Herndon et al. 2013)

“If we cannot even reproduce the original results using the same publicly available data, there is no need for further commentary.” (Miller et al, 2001)
How original authors often respond

“less realistic”, “inconsistent with the substantive literature,” and “of limited utility” (Mansfield, Milner, and Rosendorff 2002)

“fundamentally flawed”
(Peffley, Knigge, and Hurwitz 2001)

“statistical, computational, and reporting errors that invalidate its conclusions” (Gerber and Green 2005:301).
Publishing a replication study

• Good replication studies get published

• Write a **solid** paper (puzzle, relevance, hypothesis, research design, findings, discussion) – as if it was an **original** piece.

• In some fields (politics): **Don’t sell it as a replication** paper
Our estimation approach builds off of the methodology and data used by Gomez et al. (2007) ..., adding measures of electoral closeness in order to focus on how the randomly assigned cost (rain) has a different impact depending on the electoral environment.
How do domestic political institutions affect the outcomes of international trade negotiations? Specifically, are the aggregate trade barriers agreed upon by a democratic pair lower than those by a pair composed of a democracy and an autocracy? I revisit these important questions by highlighting some problematic aspect of the analysis by Mansfield, Milner, and Rosendorff (2000).

Contrary to their central conclusion, I find that whether the aggregate trade barriers are lower for a democratic pair than those for a mixed pair depends on the preferences of the decision makers involved.
Journals Open to Replication (selection)

Political Science
- The Journal of Conflict Resolution
- Research & Politics
- Political Research Quarterly

Psychology
- Psychological Science
- Perspectives on Psychological Science
- Social Psychology

Economics
- Empirical Economics
- Public Finance Review
- The American Economic Review

*original study was published in the same journal
+ home of the original ‘Many Labs’ project
# special issue dedicated to replications (March 2015)
^this journal invites replication studies
Replications by Early Career Researchers

Questioning the Effect of Nuclear Weapons on Conflict

Mark S. Bell1 and Nicholas L. Miller1

Abstract
We examine the effect of nuclear weapons on interstate conflict. Using more appropriate methodologies than have previously been used, we find that dyads in which both states possess nuclear weapons are not significantly less likely to fight wars, nor are they significantly more or less belligerent at low levels of conflict. This stands in contrast to previous work, which suggests nuclear dyads are some 2.7 million times less likely to fight wars. We additionally find that dyads in which one state possesses nuclear weapons are more prone to low-level conflict (but not more prone to war). The findings appear to be robust as nuclear-armed states expand their interest after nuclear acquisition rather than because nuclear weapons provide a shield behind which states can agress against more powerful conventional-armed states. This calls into question conventional wisdom on the impact of nuclear weapons and has policy implications for the impact of nuclear proliferation.

Near-wins and near-losses in gambling: A behavioral and facial EMG study

Thomas Herndon, Michael Ash and Robert Pollin

Externalities, Theoretical Risk, Wages, Exports, demand, Near losses, Gambling

Does High Public Debt Consistently Stifle Economic Growth?
A Critique of Reinhart and Rogoff

Near-wins and near-losses in gambling: A behavioral and facial EMG study

Thomas Herndon, Michael Ash and Robert Pollin

Abstract
This paper investigates two main claims of Reinhart and Rogoff (R&R) (2010). First, R&R argue that public debt has a consistent negative effect on economic growth. In their seminal study, R&R use a dataset of 169 countries, 1960-2005, which they claim is the largest and most representative dataset in the literature. Second, we critique R&R’s analysis and measurement of the effects of countries. We find that their data is not representative of some specific processing, despite their claims to the contrary.

WORKING PAPERS
ECOLOGICALECONOMICS
AND ECONOMICHISTORY

Working Paper No. 20 – 2014: CAN INFLATION EXPECTATIONS BE MEASURED USING COMMODITY FUTURES PRICES?

Rasheed Salehuddin (corresponding author: rksam@cam.ac.uk) and D'Maris Coffman

Centre for Financial History, University of Cambridge

Irregularities in LaCour (2014)

David Bronkam, Assistant Professor, Stanford GSB (as of July 1)
davidbronkam@stanford.edu

Jordie Knox, Graduate Student, UC Berkeley, jordieknox@berkeley.edu
Peter Aven, Assistant Professor, Yale University, peter.aven@yale.edu
May 29, 2015

Summary
We report a number of shortcomings in the replication dataset posted for LaCour and Green (2014), “When contrast change matters: An experiment on responses of expert opinion for gender equality.”[1] Using LaCour and Green’s (2014) own data and methodology, we find that the sample of experts from their study is heavily overrepresented with experts from the humanities and social sciences. We also ﬁnd that the experts in our replication study are less gender equalitarian than the experts in LaCour and Green’s (2014) study. In addition, we ﬁnd that the experts’ gender equality ratings are highly correlated with their personal gender egalitarianism. Finally, we ﬁnd that the response rates to the replication study are signiﬁcantly lower than those reported in LaCour and Green (2014). We conclude that the replication study was not conducted in a replicable manner.

Timeline of Disclosure
January – April 2015. Bronkam and Knox were surprised by LaCour and Green (2014) and wanted to extend the article’s methodology and evaluate their conclusions. We began to plan an extension. We sought to factor out the effects of gender heterogeneity based on the parent in the original data, as in the replication that they presented. On the other hand, when the paper was found, LaCour (2015). As we examined the article’s data in planning our study, it became clear that the replication study was not replicated. We therefore decided to not extend the article’s findings. Instead, we decided to contribute to the replication literature. The replication study by LaCour and Green (2014) was conducted in a replicable manner. The replication study was significantly higher than the reported in LaCour and Green (2014). We concluded that the replication study was not conducted in a replicable manner. We aimed to extend the article’s findings. Instead, we decided to contribute to the replication literature.

Journal of Experimental Political Science 1(2014) 139–148
doi:10.1017/1ps.2014.4

Information Spillovers: Another Look at Experimental Estimates of Legislator Responsiveness

Alexander Coppock

Abstract
A field experiment carried out by Border and Nickel (Border, D. M., and Nickel, D. W. 2011) on learning constituency opinions affects how legislators vote. Results from a field experiment have the potential to shed light on the determinants of legislator responsiveness. Using a difference-in-differences design, we study the effect of learning constituency opinions on legislator behavior. We find that learning constituency opinions has a significant impact on how legislators vote. We also find that the effect of learning constituency opinions is moderated by the level of uncertainty in the legislature.

INTRODUCTION
Border and Nickel (2011) report the results of an innovative field experiment testing the responsiveness of legislators to public opinion in New Mexico. Previous studies of responsiveness have focused on the relationship between public opinion and legislators’ voting behavior. This study, however, is unique in that it provides a direct test of the role of public opinion in legislators’ voting decisions. By randomly assigning some legislators to a treatment condition in which they are exposed to public opinion information, we can directly assess the impact of public opinion on legislative behavior.

The authors are grateful to Donald P. Green, Robert Elkins, Ginger Warren, Peter Arias, Carol Daniel, Cheryl Jang, and two anonymous reviewers for helpful comments and suggestions, and to Donald Green and David Nickel for providing replication materials.

© The Experimental Research Section of the American Political Science Association 2015.
Bringing the Gold Standard into the Classroom: Replication in University Teaching

NICOLE JANZ
University of Cambridge

Reproducibility is held to be the gold standard for scientific research. The credibility of published work depends on being able to replicate the results. However, there are few incentives to conduct replication studies in political science. Replications are difficult to conduct, time-consuming, and hard to publish because of a presumed lack of originality. This article sees a solution in a profound change in graduate teaching. Universities should introduce replications as class assignments in methods training or invest in new stand-alone replication workshops to establish a culture of replication and reproducibility. This article will
How to work transparently
Working reproducibly

Starting out...
- Plan file structure
- Never touch raw data!

Analysis
- Comment your code
- Keep a log of decisions

Writing up
- clarity in methodology section; appendix

Before you submit: Replicate your results !!!
What to share - quantitative

1. Readme file
2. Dataset
3. Software commands
4. Information to reconstruct data
Qualitative Analysis

- discourse analysis
- content analysis
- ethno-graphy
- interviews
- focus groups
- participant observ.
- process tracing
GUIDELINES FOR PREPARING REPPLICATION FILES

Version 2.1, May 19, 2016

William G. Jacoby
Robert N. Lupton
Michigan State University

The American Journal of Political Science requires the authors of all accepted manuscripts to provide replication files before the article enters the production stage of the publication process. The replication files for each article must be made available as a Dataset (i.e., a collection of files) located in the AJPS Dataverse on the Harvard Dataverse Network. Instructions for getting started on the AJPS Dataverse can be found in the “Quick Reference for Uploading Replication Files,”
Support your claims - qualitative

satellite images, interview transcripts, personal diaries, video clips, newspaper articles, speeches...

Transparency Appendix:

• how you evaluated persuasiveness & consistency of evidence
• logic and steps in process tracing
• Upload files and fragments e.g. partial transcripts (100-150 w.)
When to protect the data

- Confidential / proprietary data
- Protect individuals
- Informed consent obtained?

- Anonymization
- Justify why you withhold data
What’s in it for me?
1. Avoid disaster

2. Easier to write up

3. Persuade reviewers

4. Enables continuity

5. Build your reputation

Get in Touch

Twitter: @PolSciReplicate

http://PoliticalScienceReplication.wordpress.com/

Vising scholar 2016 and Catalyst

Ambassador
Useful resources


Materials – Transparent Workflow


• Open Science Framework. Transparency and Openness Promotion (TOP) Guidelines. [https://cos.io/top/](https://cos.io/top/)

• TIER Documentation Protocol [https://www.haverford.edu/project-tier/protocol-v2](https://www.haverford.edu/project-tier/protocol-v2)
Adding value to a replication

1. Theoretical contribution: questioning the arguments
2. Statistical contribution

<table>
<thead>
<tr>
<th>Sample size:</th>
<th>Model specification:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power calculations (how big should the sample be?)</td>
<td>Standard errors treatment, LDV, lags</td>
</tr>
<tr>
<td>More years, more countries (units)</td>
<td>Interactions</td>
</tr>
<tr>
<td>New samples (experiments)</td>
<td>Dummy variables</td>
</tr>
<tr>
<td>Different subsets of your data set (e.g. only OECD countries)</td>
<td>Omitted variables</td>
</tr>
<tr>
<td>Missing data handling (multiple imputation)</td>
<td>Reversed causality</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Changing measurements:</th>
<th>Robustness/Sensitivity checks:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change of variables: %GDP, log transformation, different ways of dealing with negative values for logging, different measurement for the same variable</td>
<td>How much do betas and standard errors change when we change model specifications? Are they very ‘sensitive’ even to small changes/outliers?</td>
</tr>
</tbody>
</table>