

Computational Reproducibility: Expectations, Challenges and Recommendations

JEDI 2022 Workshop – *May the force be with you: Resources to help journal editors advance their fields* May 20, 2022

Ana Trisovic, Harvard Biostatistics

Review board member of the Journal of Systems Research (JSys) Early-career board member of Harvard Data Science Review (HDSR)

"An article about computational science in a scientific publication is **not** the scholarship itself, it is merely **advertising** of the scholarship. The actual scholarship is the software, [data] ... and set of instructions which generated the figures. "

~ Prof Claerbout





 Replication dataset - a bundle of data, code and other files needed to reproduce a published study



Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse >

Replication Data for: How Political Parties Shape Public Opinion in the Real World

The second s				
ter inter-	Bisgaard, Martin; Rune Slothuus, 2020, "Replication Data for: How Political Parties Shape Public Opinion in the Real World", https://doi.org/10.7910/DVIV/Z5BTCO, Harvard Dataverse, V2, UNF:6:'TJYX+kJtxsSZUNEND/3GGg== [fileUNF]		Access Dataset -	
			Contact Owner	Share
	Cite Dataset +	Learn about Data Citation Standards.	Dataset Metrics 🕄	
Name of			1,092 Downloads 🕄	
Description 😔		How powerful are political parties in shaping citizens' opinions? Despite longstanding interest in the flow of influence between partisan ellets and citizens, few studies to date examine how citizens react when their party changes its position on a major issue in the real word. We present a rare quasi-experimental panel study of how citizens responded when their political party studdenly reversed its position on two major and salient welfare issues in Denmark. With a five-wave panel survey collected just around these two events, we show that citizens' policy opinions changed immediately and substantially when their party switched its policy positioneven when the new position went against citizens' previously held views. These findings advance the current, largely experimental literature on partisan elite influence. (2020-03-26)		
Subject 🕄		Social Sciences		
Keyword 🕢		Party cues, Political parties, Elite influence, Motivated reasoning, Polarization, Public opinion, Panel survey		
Related Publication 🕢		Bisgaard, Martin, and Rune Slothuus. [date]. "How Political Parties Shape Public Opinion in the Real World." American Journal of Political Science Forthcoming. http://ajps.org/		
Notes 9		This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill. The associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the Center for Open Science.		
Files Meta	adata Terms	Versions		
Search this d	ataset	4		
Filter by File Type: All + Access: All +				11 Sort -
1 to 10 c	of 25 Files			Ł Download -
	build_data.R R Syntax - 12.1 Kl Published Jun 29, 56 Downloads MD5: a94597	B 2020		● ₹.
- 1	codebook ess. Adobe PDF - 508. Published Jun 29, 46 Downloads	pdf 8 KB 2020		. ∎.

www.nature.com/scientificdata

scientific data

Check for updates

OPEN A large-scale study on research ANALYSIS code quality and execution

Ana Trisovic¹², Matthew K. Lau², Thomas Pasquier³ & Mercè Crosas¹

This article presents a study on the quality and execution of research code from publicly-available replication datasets at the Harvard Dataverse repository. Research code is typically created by a group of scientists and published together with academic papers to facilitate research transparency and reproducibility. For this study, we define ten questions to address aspects impacting research reproducibility. For this study, we define ten questions to address aspects impacting research reproducibility and reuse. First, we retrieve and analyze more than 2000 replication datasets with over 9000 unique R files published from 2010 to 2020. Second, we execute the code in a clean runtime environment to assess its ease of reuse. Common coding errors were identified, and some of them were solved with automatic code cleaning to aid code execution. We find that 74% of R files failed to complete without error in the initial execution, while 56% failed when code cleaning was applied, showing that many errors can be prevented with good coding practices. We also analyze the replication datasets from journals' collections and discuss the impact of the journal policy strictness on the code re-execution rate. Finally, based on our results, we propose a set of recommendations for code dissemination aimed at researchers, journals, and repositories.

Introduction

Researchers increasingly publish their data and code to enable scientific transparency, reproducibility, reuse, or compliance with funding bodies, journals, and academic institutions¹. Reusing data and code should propel new research and save researchers' time, but in practice, it is often easier to write new code than reuse old. Even attempting to reproduce previously published results using the same input data, computational steps, methods, and code has shown to be troublesome. Studies have reported a lack of research reproducibility^{2,3} often caused by inadequate documentation, errors in the code, or missing files.

- Dataverse = a free and open-source software platform to archive, share, and cite research data
 - Focus on making research data and code available
- Retrieved 2109 publicly available replication datasets containing 9078 R files
- Over 94% of the datasets belonged to social sciences

The **Dataverse** Project

scientific data

Check for updates

www.nature.com/scientificdata

OPEN A large-scale study on research ANALYSIS code quality and execution

Ana Trisovic ¹^M, Matthew K. Lau ², Thomas Pasquier ³ & Mercè Crosas¹

This article presents a study on the quality and execution of research code from publicly-available replication datasets at the Harvard Dataverse repository. Research code is typically created by a group of scientists and published together with a cademic papers to facilitate research transparency and reproducibility. For this study, we define ten questions to address aspects impacting research reproducibility. For this study, we define ten questions to address aspects impacting research reproducibility and reuse. First, we retrieve and analyze more than 2000 replication datasets with over 9000 unique R files published from 2010 to 2020. Second, we execute the code in a clean runtime environment to assess its ease of reuse. Common coding errors were identified, and some of them were solved with automatic code cleaning to aid code execution. We find that 74% of R files failed to complete without error in the initial execution, while 56% failed when code cleaning was applied, showing that many errors can be prevented with good coding practices. We also analyze the replication datasets from journals' collections and discuss the impact of the journal policy strictness on the code re-execution rate. Finally, based on our results, we propose a set of recommendations for code dissemination aimed at researchers, journals, and repositories.

Introduction

Researchers increasingly publish their data and code to enable scientific transparency, reproducibility, reuse, or compliance with funding bodies, journals, and academic institutions¹. Reusing data and code should propel new research and save researchers' time, but in practice, it is often easier to write new code than reuse old. Even attempting to reproduce previously published results using the same input data, computational steps, methods, and code has shown to be troublesome. Studies have reported a lack of research reproducibility^{2,3} often caused by inadequate documentation, errors in the code, or missing files.

Our data collection workflow

1. Replication dataset is retrieved from Harvard Dataverse to AWS



Our data collection workflow

- 1. Replication dataset is retrieved from Harvard Dataverse to AWS
- 2. We collect data on the content, install used libraries and attempt automatic code re-execution



Our data collection workflow

- 1. Replication dataset is retrieved from Harvard Dataverse to AWS
- 2. We collect data on the content, install used libraries and attempt automatic code re-execution
- The re-execution result and other collected data are passed to the backend database for analysis









Portion of replication datasets with re-executable code files



Portion of replication datasets with re-executable code files





Journal average: 47% Total average: 45%

What can journals do?



1. Verification or review of deposited materials



- 1. Verification or review of deposited data and code
- 2. Reproducibility checklist or README templates for authors

- 1. Verification or review of deposited data and code
- 2. Reproducibility checklist or README templates for authors



A template README for social science replication packages.

The template README provided on this website is in a form that follows best practices as defined by a number of data editors at social science journals.

Template README and Guidance

INSTRUCTIONS: This README suggests structure and content that have been approved by various journals, see Endorsers. It is available as Markdown/txt, Word, LaTeX, and PDF. In practice, there are many variations and complications, and authors should feel free to adapt to their needs. All instructions can (should) be removed from the final README (in Markdown, remove lines starting with > INSTRUCTIONS). Please ensure that a PDF is submitted in addition to the chosen native format.

Overview

INSTRUCTIONS: The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper. Start by providing a brief overview of the available material and a brief guide as to how to proceed from beginning to end.

Example: The code in this replication package constructs the analysis file from the three data sources (Ruggles et al, 2018; Inglehart et al, 2019; BEA, 2016) using Stata and Julia. Two main files run all of the code to generate the data for the 15 figures and 3 tables in the paper. The replicator should expect the code to run for about 14 hours.

https://social-science-data-editors.github.io

Data Availability and Provenance Statements

@larsvil

- 1. Verification or review of deposited data and code
- 2. Reproducibility checklist or README templates for authors
- 3. Integration with data and software repositories



- 1. Verification or review of deposited data and code
- 2. Reproducibility checklist or README templates for authors
- 3. Integration with data and software repositories

to code, for any arXiv paper ode, for any arXiv paper	D D S	tHub Action ataverse Upl v1.0 (Latest version)	oader Action
Papers With Code	Dataverse	e Uploader	
	This action uploads the repository content to a Dataverse dataset.		
partners with	Input parameters		
arXiv org	To use this action,	you will need the follow	ing input parameters: Description
	DATAVERSE_TOKE	N Yes	This is your personal acc token that you can create your Dataverse instance the Dataverse guide). Sa your token as a secret variable called
	o code, for any arXiv paper ode, for any arXiv paper Papers With Code partners with arXiv.org	so code, for any arXiv paper ode, for any arXiv paper (IIII) Papers With Code partners with arXiv.org Dataverse This action upload Input parametr To use this action, Parametr	o code, for any arXiv paper ode, for any arXiv paper papers With Code partners with arXiv.org GitHub Action Dataverse Uploader This action uploads the repository content Input parameters To use this action, you will need the follow Parameter Required DATAVERSE_TOKEN Yes

Email: anatrisovic@g.harvard.edu Twitter & GitHub: atrisovic





- 1. **Verification or review** of deposited materials
- 2. Reproducibility **checklist** or README **templates** for authors
- 3. **Integration** with data and software repositories