Computational reproducibility: Examining verification errors and frictions

Cheryl A. Thompson & Thu-Mai L. Christian

H. W. Odum Institute for Research in Social Science,
University of North Carolina at Chapel Hill

> **Computational reproducibility** refers to changes in scientific practice and reporting standards to accommodate the use of computational technology...in particular whether the **same results can be obtained from the data and code used in the original study.** (Stodden, 2015)

Computational reproducibility =

Transparency + Reproducibility of computation

(NASEM, 2019)

AJPS Verification Policy

AJPS

VOLUME 63 NUMBER 3 JULY 2019

AMERICAN JOURNAL of POLITICAL SCIENCE

Committed to significant advances in knowledge and understanding of politics, governance, and citizenship, and to the public value of political science research

Official Publication of the
Midwest Political Science Association
www.mpsanet.org • www.ajps.org

- The corresponding author of a manuscript that is accepted for publication in the *American Journal of Political Science* **must provide materials that are sufficient to enable interested researchers to verify all of the analytic results that are reported in the text and supporting materials.**
- When the final draft of the manuscript is submitted, the materials will be **verified to confirm that they do, in fact, reproduce the analytic results reported in the article.**
- **Publication in the *American Journal of Political Science* is contingent** upon provision of complete verification materials and **successful verification**…

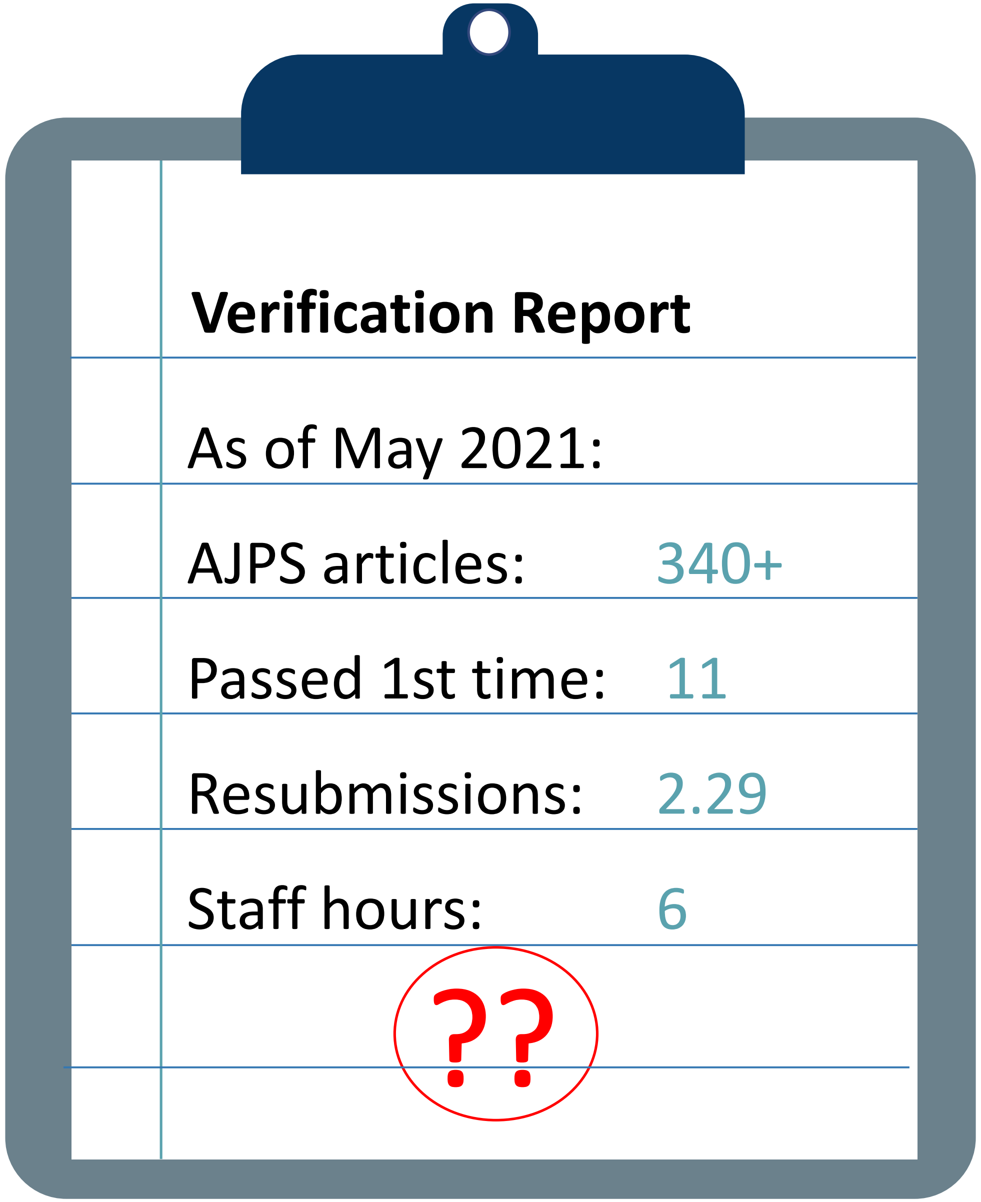
(AJPS Verification Policy, n.d., https://ajps.org/ajps-verification-policy)

Curation

- ✓ Review replication package for completeness
- ✓ Identify confidentiality / copyright issues
- ✓ Identify incomplete, inconsistent, or missing variable and value labels
- ✓ Enhance descriptive metadata
- ✓ Assess file formats for suitability for long-term preservation

Verification

- ✓ Review code for inclusion of commands and comments required to re-produce reported results
- ✓ Compile and execute code
- ✓ Identify errors in non-executable code
- ✓ Compare outputs to tables, figures, and other reported results in the manuscript



Verification Report

As of May 2021:

AJPS articles:	340+
Passed 1st time:	11
Resubmissions:	2.29
Staff hours:	6

??

Qualitative Study

- RQ: What are the challenges that authors face in complying with computational reproducibility and verification policies?

- Sample of 105 manuscripts (2017-2019)
  - Verification report: dates, result, open data, curation notes, verification notes, resubmissions

- Qualitative coding and analysis

Sample

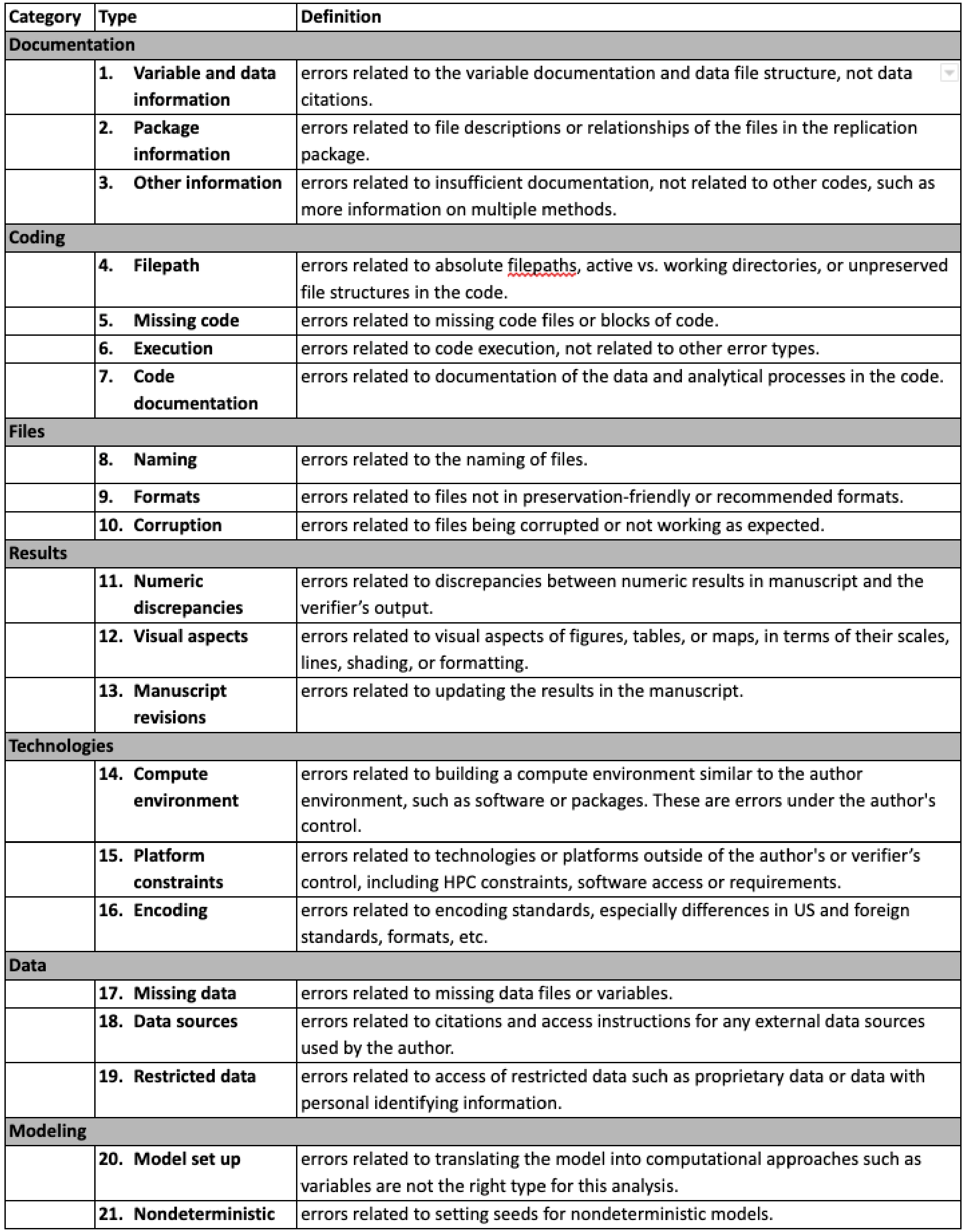
- Verification characteristics:
    - Mean number of resubmissions: 2.4
    - Verified on initial submission: 1.9% (2)

- Package characteristics at initial submission:
    - Mean number of files: 10.9
    - Mean number of lines of code: 2,667.8
    - Mean number of programming languages: 1.7

Typology of Verification Errors

21 error types -> 7 categories

- Documentation
- Coding
- Files
- Technologies
- Data
- Modeling
- Results



Category	Type	Definition
Documentation		
	1. Variable and data information	errors related to the variable documentation and data file structure, not data citations.
	2. Package information	errors related to file descriptions or relationships of the files in the replication package.
	3. Other information	errors related to insufficient documentation, not related to other codes, such as more information on multiple methods.
Coding		
	4. Filepath	errors related to absolute filepaths, active vs. working directories, or unpreserved file structures in the code.
	5. Missing code	errors related to missing code files or blocks of code.
	6. Execution	errors related to code execution, not related to other error types.
	7. Code documentation	errors related to documentation of the data and analytical processes in the code.
Files		
	8. Naming	errors related to the naming of files.
	9. Formats	errors related to files not in preservation-friendly or recommended formats.
	10. Corruption	errors related to files being corrupted or not working as expected.
Results		
	11. Numeric discrepancies	errors related to discrepancies between numeric results in manuscript and the verifier’s output.
	12. Visual aspects	errors related to visual aspects of figures, tables, or maps, in terms of their scales, lines, shading, or formatting.
	13. Manuscript revisions	errors related to updating the results in the manuscript.
Technologies		
	14. Compute environment	errors related to building a compute environment similar to the author environment, such as software or packages. These are errors under the author's control.
	15. Platform constraints	errors related to technologies or platforms outside of the author's or verifier’s control, including HPC constraints, software access or requirements.
	16. Encoding	errors related to encoding standards, especially differences in US and foreign standards, formats, etc.
Data		
	17. Missing data	errors related to missing data files or variables.
	18. Data sources	errors related to citations and access instructions for any external data sources used by the author.
	19. Restricted data	errors related to access of restricted data such as proprietary data or data with personal identifying information.
Modeling		
	20. Model set up	errors related to translating the model into computational approaches such as variables are not the right type for this analysis.
	21. Nondeterministic	errors related to setting seeds for nondeterministic models.

Documentation errors

Type	Description
1. Variable and data	errors related to the variable documentation and data file structure, not data citations.
2. Package	errors related to file descriptions or relationships of the files in the replication package.
3. Other information	errors related to insufficient documentation, not related to other codes, such as more information on multiple methods.

Coding errors

Type	Description
1. Filepath	errors related to absolute filepaths, active vs. working directories, or unpreserved file structures in the code.
2. Missing code	errors related to missing code files or blocks of code.
3. Execution	errors related to code execution, not related to other error types.
4. Code documentation	errors related to documentation of the data and analytical processes in the code.

Technologies errors

Type	Description
1. Compute environment	errors related to building a compute environment similar to the author environment, such as software or packages. These are errors under the author's control.
2. Platform constraints	errors related to technologies or platforms outside of the author's or verifier’s control, including HPC constraints, software access or requirements.
3. Encoding	errors related to encoding standards, especially differences in US and foreign standards, formats, etc.

Frictions

Data sharing

Data as knowledge production vs. data as a final product

Code sharing

Code as a means vs. code as a final product

Policy and standards

Formal policies standards vs. informal research practices

Temporal

Project lifecycle vs. verification at end

Technology

Proliferation of RR tools vs. researcher tool preferences

Labor

It's nobody's job

Next steps

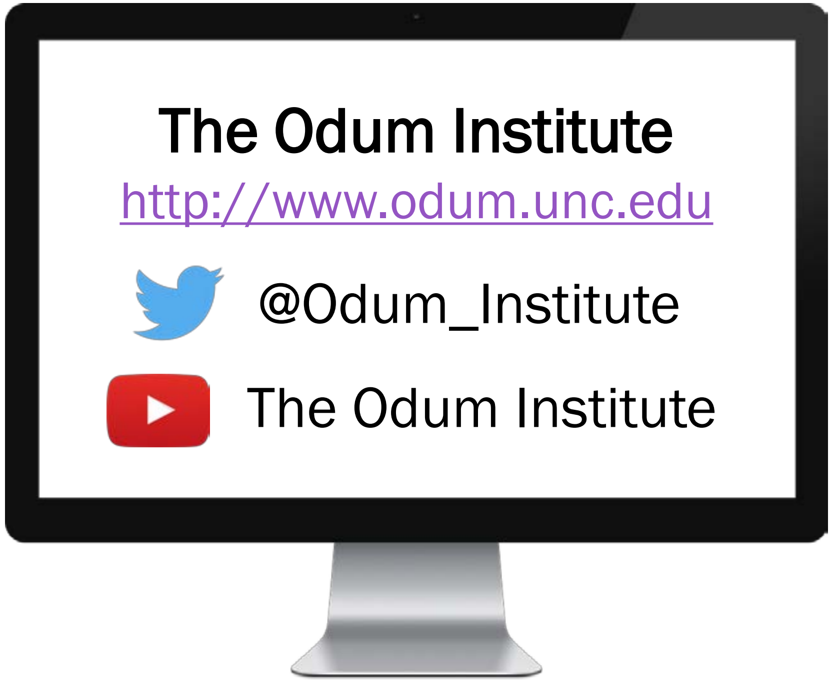
- Future analysis:
  - model to understand which errors lead to longer verification times

- Targeted guidance for authors

- Inform computational reproducibility trainings

Connect with the Odum Institute

Cheryl A. Thompson
cathompson@unc.edu



The Odum Institute
http://www.odum.unc.edu

@Odum_Institute

The Odum Institute

Sample

- Author characteristics:
  - Non-US-based corresponding author: 26.7% (28)
  - Corresponding author had participated previously in AJPS verification: 11.4% (12)

Data errors

Type	Description
1. Missing data	errors related to missing data files or variables.
2. Data sources	errors related to citations and access instructions for any external data sources used by the author.
3. Restricted data	errors related to access of restricted data such as proprietary data or data with personal identifying information.