

Analysis of Verified Replication Packages from the American Journal for Political Science

Mentor: Craig Willis

Summer Intern: Jingxian Na

INTRODUCTION

- Since 2015, the American Journal for Political Science (AJPS) started to require researchers upload replication materials (code, data, and metadata) that's necessary to reproduce the analytic results and cooperated with Odum institute on the verification process:

The screenshot displays the AJPS website with a green and blue header. The main navigation bar includes links for Home, About, Contact, Subscribe/Join, and Current Issue. Below this, a secondary navigation bar lists Home, AJPS Articles, Reviews, Manuscripts, Data Integrity, and About AJPS. The main content area features the 'AJPS Verification Policy' section, which explains the requirements for authors to provide replication materials. To the right, there is a 'TRANSLATE' section with a language selection dropdown, a search bar, and a 'FOLLOW AJPS VIA EMAIL' section with an email subscription form. At the bottom right, the 'Current Issue' is listed as July 2019.

Home About Contact Subscribe/Join Current Issue

AJPS

AMERICAN JOURNAL of POLITICAL SCIENCE

HOME AJPS ARTICLES REVIEWS MANUSCRIPTS DATA INTEGRITY ABOUT AJPS

AJPS Verification Policy

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. The document titled "*American Journal of Political Science Guidelines for Preparing Replication Files*"* provides useful information about what information is needed and how it should be organized. All verification files must be stored in a Dataset within the *AJPS* Dataverse, on the Harvard Dataverse Network. Note that authors also can make their verification files available elsewhere (e.g., their personal website, other data repositories, etc.) as long as all of the necessary files are included in the Dataset on the *AJPS* Dataverse.

The corresponding author should prepare and upload verification materials to the *AJPS* Dataverse before submitting the final draft of the accepted manuscript. The "*American Journal of Political Science Quick Reference for Uploading Replication Files*"* provides information about creating a Dataset on the *AJPS* Dataverse and depositing materials into it. The "*American Journal of Political Science Quantitative Data Verification Checklist*" and "*American Journal of Political Science Qualitative Data Verification Checklist*" are designed to help authors ensure that they have provided all necessary files.

TRANSLATE

Select Language

Powered by Google Translate

Search this website...

FOLLOW AJPS VIA EMAIL

Enter your email address to receive notifications of new AJPS Author Summaries and posts from the editors.

Email Address

Current Issue – July 2019

The *American Journal of Political Science* (AJPS) is the flagship journal of the

INTRODUCTION



- Therefore, it should be more likely to successfully reproduce the analytic results from packages deposited after the implementation of this policy than from packages deposited before the policy.
- Used python to characterize datasets in AJPS Dataverse, including plot generation and identification of verified datasets.
- Then tried to replicate two datasets that contains only R scripts.

First Step: download and characterize

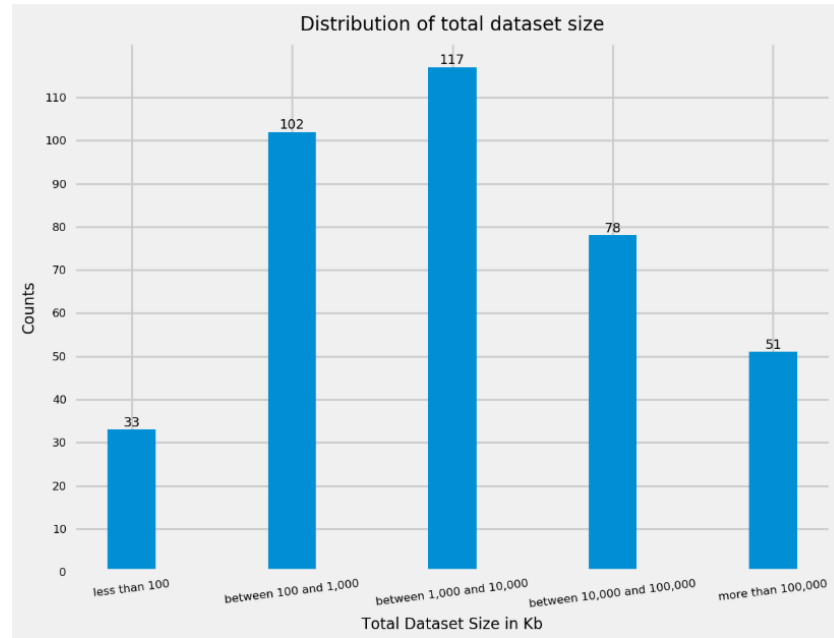
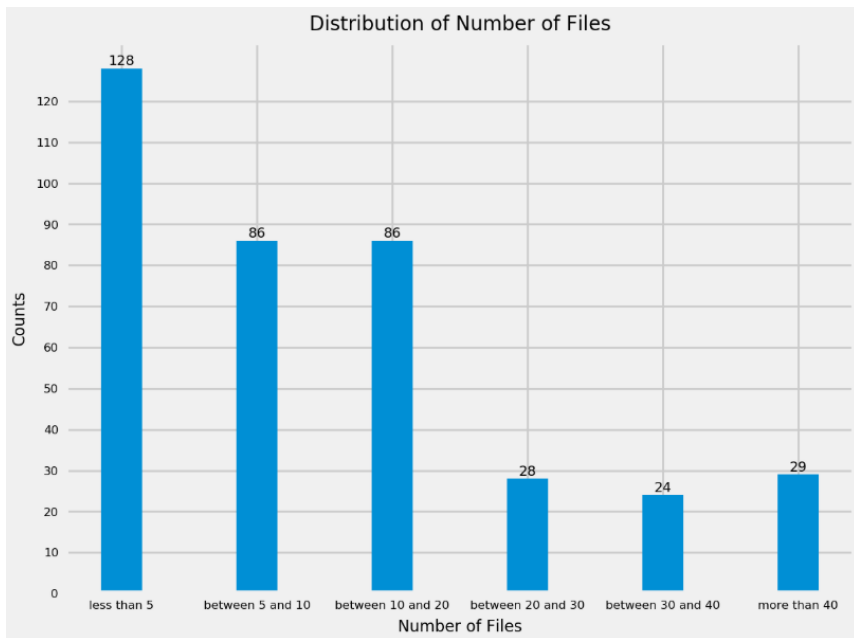


Using Python to:

- Get a list of DOI of all existing datasets in AJPS dataverse through its search API
- Get the metadata of all these datasets through OAI_ORE exporter using DOI list
- Do some analysis on the metadata

Second Step: generate some plots

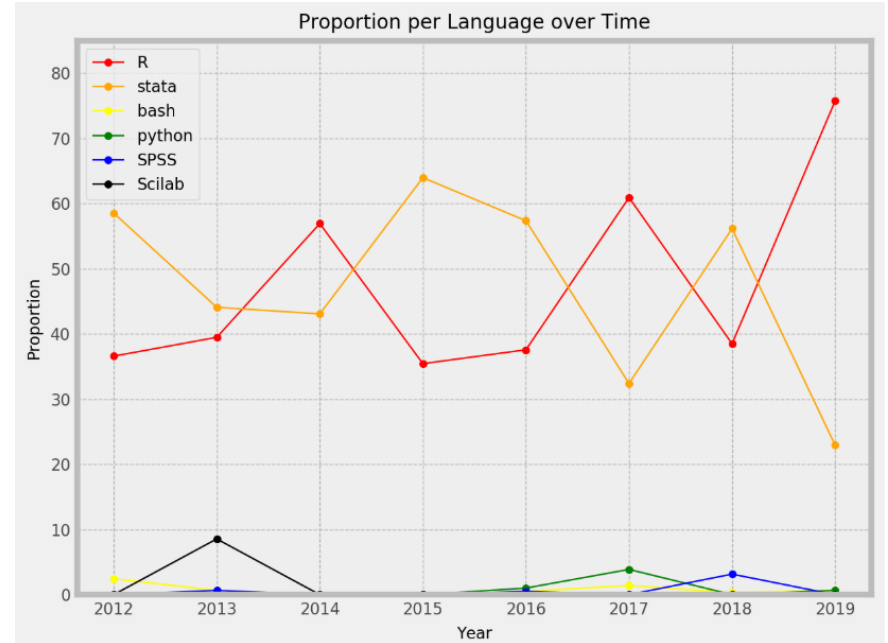
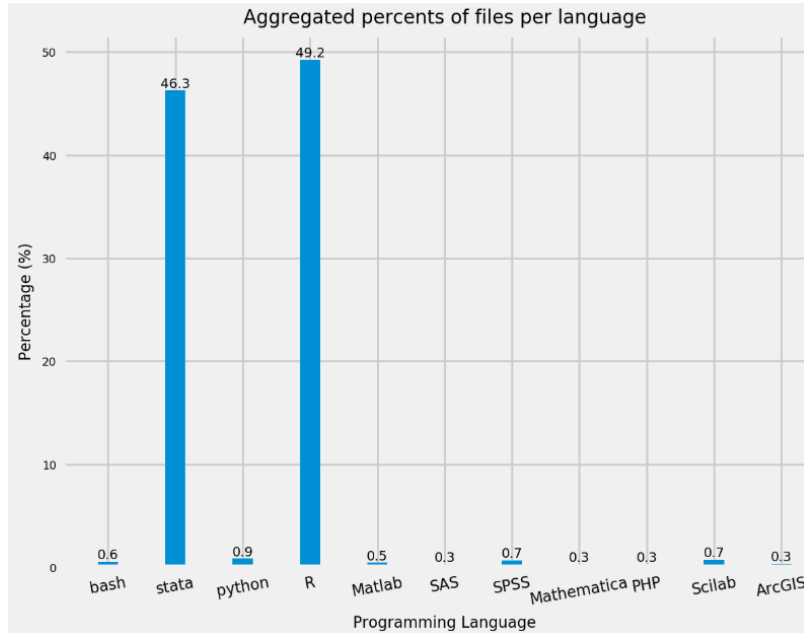
Using the csv file I got from the first step, I drew some plots to get a whole view of the AJPS dataverse, notebook link: <https://dashboard.wholetale.org/tale/view/5d406051cc895d0f74354993>



Second Step: generate some plots

Using the csv file I got from the first step, I drew some plots to get a whole view of the AJPS

Dataverse:



Second Step: generate some plots



- Based on the histograms and plots I got from the second step, R and stata are the two most popular coding language in AJPS Dataverse.
- As a result, I decided to compare the verification process of two datasets that contains only R files, one of them has gone through the Odum verification and the other has not.

Third Step: identify verified datasets

As mentioned above, the replication policy was adopted by AJPS on March 2015. However, there are also some datasets in AJPS Dataverse that was published later than that date and have not been checked by Odum:

 **Replication Data for: Influencing the Bureaucracy: The Irony of Congressional Oversight** Version 1.1

Clinton, Joshua, 2018, "Replication Data for: Influencing the Bureaucracy: The Irony of Congressional Oversight", <https://doi.org/10.7910/DVN/QC0RSU>, Harvard Dataverse, V1, UNF:6:KwhNNUJWtbkS1sJ97/BzQ== [fileUNF]

[Cite Dataset](#) 

[Learn about Data Citation Standards.](#)

Description 

Does the President or Congress have more influence over policy making by the bureaucracy? Despite a wealth of theoretical guidance, progress on this important question has proven elusive due to competing theoretical predictions and severe difficulties in measuring agency influence and oversight. We use a survey of federal executives to assess political influence, congressional oversight and the policy preferences of agencies, committees, and the president on a comparable scale. Analyzing variation in political influence across and within agencies reveals that Congress is less influential relative to the White House when more committees are involved. While increasing the number of involved committees may maximize the electoral benefits for members, it may also undercut the ability of Congress as an institution to collectively respond to the actions of the presidency or the bureaucracy. (2014)

Subject 

Social Sciences

Keyword 

Congress, Bureaucracy, Agencies, President, Committees, Oversight

Related Publication 

Clinton, Joshua. 2014. "Influencing the Bureaucracy: The Irony of Congressional Oversight." *American Journal of Political Science*. 58 (2): 387-401. doi: 10.1111/ajps.12066

[Files](#)

[Metadata](#)

[Terms](#)

[Versions](#)

 [Export Metadata](#) 

Citation Metadata 

Dataset Persistent ID 

doi:10.7910/DVN/QC0RSU

Publication Date 

2018-12-04

Third Step: identify verified datasets

- In most cases, verified datasets would have corresponding statement in its “Notes” in Metadata, like “This dataset underwent an independent verification process...”.
- However, these statements always varies from dataset to dataset:

```
unique_note,count
There is no note in this dataset.,98
"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. \r\n<br></br>\r\nThe associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the <a href=\"https://osf.io/tvyxz/wiki/home/\" target=\"_blank\">Center for Open Science</a>.<br></br>\r\n<img src=\"http://www.odum.unc.edu/archive/cos_open_data.png\" alt=\"Open Data Badge\" height=\"77\" width=\"80\">\r\n<img src=\"http://www.odum.unc.edu/archive/cos_open_materials.png\" alt=\"Open Materials Badge\" height=\"77\" width=\"80\">\",61
"This dataset underwent an independent verification process that replicated the tables and figures in the main article and those included in supplementary materials on which primary findings are contingent. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.\r\n<br></br>\r\nThe associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the <a href=\"https://osf.io/tvyxz/wiki/home/\" target=\"_blank\">Center for Open Science</a>.<br></br>\r\n<img src=\"http://www.odum.unc.edu/archive/cos_open_data.png\" alt=\"Open Data Badge\" height=\"77\" width=\"80\">\r\n<img src=\"http://www.odum.unc.edu/archive/cos_open_materials.png\" alt=\"Open Materials Badge\" height=\"77\" width=\"80\">\",29
"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. \r\n\r\n<br></br>\r\nThe associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the <a href=\"https://osf.io/tvyxz/wiki/home/\" target=\"_blank\">Center for Open Science</a>.<br></br>\r\n<img src=\"http://www.odum.unc.edu/archive/cos_open_data.png\" alt=\"Open Data Badge\" height=\"77\" width=\"80\">\r\n<img src=\"http://www.odum.unc.edu/archive/cos_open_materials.png\" alt=\"Open Materials Badge\" height=\"77\" width=\"80\">\",19
"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.\r\n\r\n<br></br>\r\nThe associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the <a href=\"https://osf.io/tvyxz/wiki/home/\" target=\"_blank\">Center for Open Science</a>.<br></br>\r\n<img src=\"http://www.odum.unc.edu/archive/cos_open_data.png\" alt=\"Open Data Badge\" height=\"77\" width=\"80\">\r\n<img src=\"http://www.odum.unc.edu/archive/cos_open_materials.png\" alt=\"Open Materials Badge\" height=\"77\" width=\"80\">\",18
```

Third Step: identify verified datasets



- After analyzing all kinds of unique notes, we found that “independent verification” is a proper phrase to pick out all verified datasets.
- Used this phrase as judging condition, divided verified datasets and others into two dataframes for further analysis.

Fourth Step: draft verification report

The first dataset I verified is “Replication Data for: Campaign Contributions Facilitate Access to Congressional Officials: A Randomized Field Experiment”, it was published before March 2015 and has not gone through the Odum verification. There are two files contained in this dataset.

The screenshot displays the RStudio environment with the following components:

- Files Panel:** Shows two files: `kalla-broockman-donor-access-2015-c...` (Plain Text - 2.0 KB - 2015-1-5 - 146 MD5: 67cca99268af7c8fe177ec11c) and `kalla-broockman-donor-access-2015-c...` (Tabular Data - 1.9 KB - 2015-1-5 - 2 4 Variables, 191 Observations - UN).
- Environment Panel:** Lists variables: `cumulative.probs` (num [1:2, 1:6] 0.8236 0.0781 0.0236 0.125 0.055...), `data` (191 obs. of 4 variables), and `perms` (Large matrix (1910000 elements, 145.7 Mb)).
- Code Editor:** Contains R code for data manipulation and analysis, including `cumulative.probs[2,] <- cumulative.probs[2,6]`, `round(cumulative.probs, digits = 3)*100`, and `library(MASS)`.
- Console:** Shows the output of the R code, including `[1] 0.05130413` and `$p.value [1] 0.25386`.
- Plots Panel:** Displays a density plot titled `density.default(x = ate.dist.under.sharp.null)` with the x-axis labeled `N = 100000 Bandwidth = 0.004135`.
- Launched Tales:** A sidebar on the right showing the dataset title and a small R logo.

Fourth Step: draft verification report

There is a table and two figures in the corresponding article:

TABLE 1 Results: Access Gained in Constituent and Revealed Donor Conditions

Level of Official Group Met	Constituent Condition Frequency	Revealed Donor Condition Frequency	Constituent Condition Cumulative Probability	Revealed Donor Condition Cumulative Probability	p-Value: Revealed Donors More Likely to Gain Access at or above This Rank
Member of Congress	2.4%	7.8%	2.4%	7.8%	p = .07
Chief of Staff	0.0%	4.7%	2.4%	12.5%	p = .006
Legislative Director or Deputy Chief of Staff	3.1%	6.2%	5.5%	18.8%	p = .005
DC-Based Legislative Assistant or Local District Director	25.2%	18.8%	30.7%	37.5%	p = .17
Other District-Based Staffer	12.6%	10.9%	43.3%	48.4%	p = .26
No Meeting	56.7%	51.6%	100%	100%	—

FIGURE 1 Access Gained to Congressional Staffers, by Experimental Condition

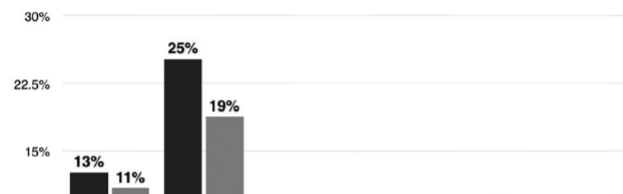
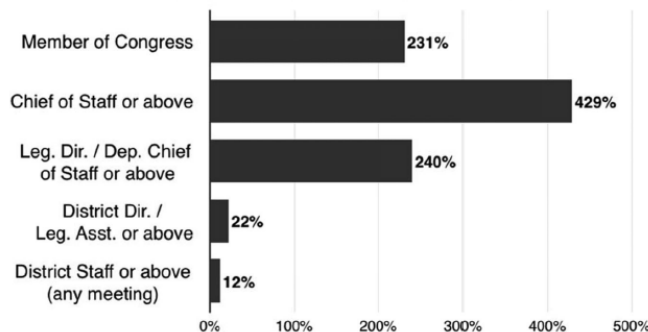


FIGURE 2 Percentage Increase in Access Revealed Donors Gained, at or above Each Level.



Note: Each bar shows the percent increase in the share of meetings that occurred at or above each level in the Revealed Donor condition relative to the Constituent condition. These can be obtained by comparing the fourth column of 1 to the third column.

Fourth Step: draft verification report

Here is the output of the R file in dataset:

```
> round(cumulative.probs, digits = 3)*100
```

```
      [,1] [,2] [,3] [,4] [,5] [,6]  
[1,]  2.4  2.4  5.5 30.7 43.3 100  
[2,]  7.8 12.5 18.8 37.5 48.4 100
```

```
[1] 5  
$ate  
[1] 0.05450295  
[1] 2  
$ate  
[1] 0.06791339
```

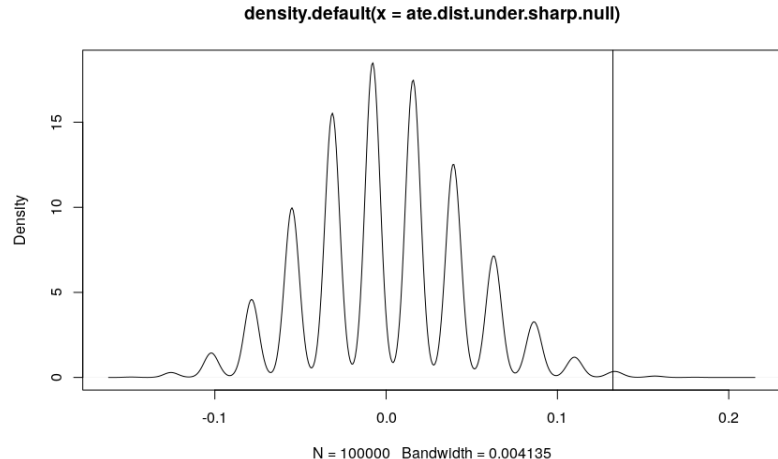
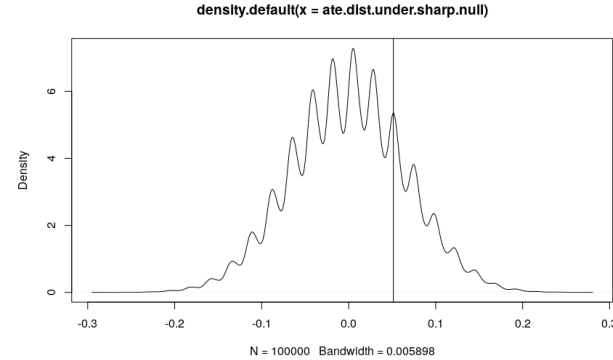
```
$p.value  
[1] 0.07332  
$p.value  
[1] 0.17004
```

```
[1] 4  
$ate  
[1] 0.101378  
[1] 1  
$ate  
[1] 0.05130413
```

```
$p.value  
[1] 0.00634
```

```
[1] 3  
$ate  
[1] 0.1323819  
$p.value  
[1] 0.25386
```

```
$p.value  
[1] 0.00506
```



Fourth Step: draft verification report

As for the verification report, these problems are all from AJPS data verification checklist:

Curation checklist

Supporting Documentation and Information:

1. Does data package include a README file (.txt format) containing the names of all files with a brief description and any other important information regarding how to replicate the findings (i.e., the order files need to be run, etc.)?

No, this package does not contain a README file.

2. Does data package include a Codebook (.pdf format) with variable definitions for all variables in the analysis dataset(s) and value labels for categorical variables?

No, this package does not contain a Codebook.

3. Does data package include clear information regarding the software version used to conduct analysis?

No, this package does not mention the specific software version.

4. Does data package include source datasets? If so, does it include complete references for source datasets?

No, the dataset used to generate plots is created by the authors of this paper.

3. Does data package include software command file(s) for reconstructing the analysis dataset from the original data source and/or extracting and merging multiple original source datasets, including information on source dataset version and access date(s)?

No, the dataset used to generate plots is created by the authors of this paper.

Command Files & Tools:

1. Does data package include commands needed to reproduce all tables, figures and other analytical results presented in the article and supplementary materials?

No, the output of the R code in the data package does not match with tables and figures in the article.

2. Does data package include commands/instructions for installing macros or packages?

Yes, the R code file contains commands for installing packages.

3. Does data package include comment statements used to explain the analysis steps and distinguish commands for tables, figures, and other outputs?

Yes, there are brief comment statements before most blocks of the R code.

Fourth Step: draft verification report

Legislative offices were then randomly assigned to treatment conditions within each of these 64 blocks, with one office in each block being randomly assigned to the Revealed Donor condition and the other two offices assigned to the Constituent condition. Treatment offices were selected by block—assigning each observation a random number with Stata 12's `runiform()` function and selecting the office in each block with the lowest random number for the treatment group.

Data curation result: Major issues

Data curation notes:

- The data package is lack of codebook and README file.

Replication result: Major issues

Replicate notes:


Code:

- The output of the R code in the data package does not match with table and figures in the article (Table 1, Figure 1 & 2). Please upload the corresponding R script (and probably stata script) to the data package.

Jackman, and Rivers 2004; Jackman 2013), the number of members of the political organization that resided within 40 miles of the district office where the meeting would be held, and Barack Obama's share of the 2012 two-party presidential vote in the district.⁶ Legislative offices were allocated to these blocks using `blockTools` in R, which seeks to construct the blocking scheme that minimizes the differences between observations within each block as determined by the Mahalanobis distance (Moore and Schnakenberg 2013).

Fourth Step: draft verification report

The second dataset I verified is “Replication Data for: Voter Buying: Shaping the Electorate through Clientelism”, it was published after March 2015 and has already gone through the Odum verification. There are four files contained in this dataset.

**Replication Data for: Voter Buying: Shaping the Electorate through Clientelism** Version 1.2

Hidalgo, F. Daniel; Nichter, Simeon, 2015, "Replication Data for: Voter Buying: Shaping the Electorate through Clientelism", <https://doi.org/10.7910/DVN/9OOLQ7>, Harvard Dataverse, V1, UNF:6:PSzaEK7iepLPu7IsbzFNiw== [fileUNF]

[Learn about Data](#)

Description

Studies of clientelism typically assume that political machines distribute rewards to persuade or mobilize the existing argue that rewards not only influence actions of the electorate, but can also shape its composition. Across the world, employ "voter buying" to import outsiders into their districts. Voter buying demonstrates how clientelism can underpin and offers an explanation of why machines deliver rewards when they cannot monitor vote choices. Our analyses su buying dramatically influences municipal elections in Brazil. A regression discontinuity design suggests that voter au undermined voter buying -- decreased the electorate by 12 percentage points and reduced the likelihood of mayoral r percentage points. Consistent with voter buying, these effects are significantly greater in municipalities with large vot where neighboring municipalities had large voter outflows. Findings are robust to an alternative research design usin dataset.

Subject

Social Sciences

Keyword

Clientelism, Machine politics, Fraud, Vote buying, Elections, Reelection, Electorate, Politics and government, Auditing Regression discontinuity design

Related Publication

Hidalgo, F. Daniel and Simeon Nichter. 2016. "Voter Buying: Shaping the Electorate through Clientelism." *American Political Science* 60 (2): 436-455. [doi: 10.1111/ajps.12214](https://doi.org/10.1111/ajps.12214)

Notes


This dataset underwent an independent verification process that replicated the tables and figures in the main article & included in supplementary materials on which primary findings are contingent. The verification process was carried o Institute for Research in Social Science at the University of North Carolina at Chapel Hill.


FilesMetadataTermsVersions


Search this dataset... Find


Filter by
File Type: All Access: All

1 to 4 of 4 Files

 **codebook.pdf**
Adobe PDF - 82.0 KB - 2015-6-9 - 114 Downloads
MD5: 117392521bcc698b542fd03176e1e4ed
Codebook

 **readme.txt**
Plain Text - 283 B - 2015-6-9 - 85 Downloads
MD5: 9152651bc510844e859635dbfa1c9553
Readme

 **replication_data.tab**
Tabular Data - 4.4 MB - 2015-6-9 - 205 Downloads
110 Variables, 5547 Observations - UNF:6:PSzaEK7iepLPu7IsbzFNiw==
Replication Code

 **voter_buying_replication_file.R**
Plain Text - 72.6 KB - 2015-6-9 - 150 Downloads
MD5: 9002c3a1705b00a46d967f870d8db21
Dataset

Fourth Step: draft verification report

There are seven graphs and three tables in the article, and almost all the R script outputs match with the analytic result in article, which is a huge improvement compared to the last dataset.

Curation checklist

Supporting Documentation and Information:

1. Does data package include a README file (.txt format) containing the names of all files with a brief description and any other important information regarding how to replicate the findings (i.e., the order files need to be run, etc.)?

Yes, this package contains a README file with brief descriptions of all files in it.

2. Does data package include a Codebook (.pdf format) with variable definitions for all variables in the analysis dataset(s) and value labels for categorical variables?

Yes, there is a Codebook in the package containing descriptions and primary sources of every variable in the analysis dataset, which is also included in the package.

3. Does data package Includes clear information regarding the software version used to conduct analysis?

Yes, the related information is included in the R code file in dataverse.

4. Does data package include source datasets? If so, does it include complete references for source datasets?

Yes, the data in the analysis dataset come from multiple sources, the authors mentioned the data sources in Codebook.

Analysis Dataset(s):

1. Does data package include the analysis dataset(s) in a file format readily accessible to the social science research community (i.e., text files, delimited files, Stata files, R files, SAS files, SPSS files, etc.)?

Yes, there is an R data file in the data package, includes the dataset used to generate the graphs and tables in the article.

2. Does data package include a unique case identifier variable linking each observation in the analysis dataset to the original data source?

Yes, the related information is included in the Codebook.

3. Does data package include software command file(s) for reconstructing the analysis dataset from the original data source and/or extracting and merging multiple original source datasets, including information on source dataset(s) version and access date(s)?

No, the dataset used to generate plots and tables is directly given to us in the dataverse.

Command Files & Tools:

1. Does data package include commands needed to reproduce all tables, figures, and other analytical results presented in the article and supplementary materials?

Yes, the output of the R code in the data package matches with the table and figures in the article.

Fourth Step: draft verification report

However, there are still some minor problems:

Codebook

"Voter Buying: Shaping the Electorate through Clientelism," *American Journal of Political Science*

F. Daniel Hidalgo and Simeon Nichter

Brazil Dataset for RDD ("data")

Variable Name	Description	Pr
above	Above the 80% discontinuity threshold (dummy variable based on "elecpopratio.pct")	Authors' Calculation
average_section_size	Average number of voters per precinct	www.tse.gov.br
branconulo.perpop04	Blank and null votes in 2004 as percentage of 2007 population	www.tse.gov.br; ww
branconulo.perpop08	Blank and null votes in 2008 as percentage of 2007 population	www.tse.gov.br; ww
chall_age	Age of challenger in 2004	www.tse.gov.br
codigo	Municipality code - IBGE (Census Bureau)	www.ipeadata.gov.l
dist.statecap	Distance from state capital (in kilometers)	www.ipeadata.gov.l
donation_diff	Difference in donations received by incumbent and 2nd place challenger in 2004	www.tse.gov.br
elecpopratio	June 2006 Electorate divided by July 2006 population [Note: These dates used by TSE for audit.]	www.tse.gov.br; dai
elecpopratio.pct	June 2006 Electorate divided by July 2006 population (percent) [Note: These dates used by TSE for audit.]	www.tse.gov.br; dai
electorate.june.2006	June 2006 Electorate [Note: Used by TSE for audit.]	www.tse.gov.br
electorate.perpop06	2006 Electorate as a percentage of 2007 population	www.tse.gov.br; ww

Data curation result: Minor issues

Data curation notes:

- The data package is lack of software command file for reconstructing the analysis dataset from the original source datasets.

Replication result: Success with modification

Replicate notes:

Table 1 & Figure 1:

- Lack of corresponding code to generate these two outputs.

Figure 4a:

- Typo on one parameter: 'xintercept', line 149
- Warning message: "Ignoring unknown parameters: degree", line 142

Figure 4b:

- Typo on one parameter: 'xintercept', line 164
- Warning message: "In comma_format(digits = 10) :
'digits' argument is deprecated, use 'accuracy' instead.FALSE"

Figure 6:

- Warning message: "Ignoring unknown parameters: degree", line 225

Wrap up



Final Outputs:

- One CSV file, 381 rows 24 columns, contains characteristic information of AJPS datasets.
- One jupyter notebook, showing histograms and plots, giving a whole view of AJPS Dataverse.
- Git repo: <https://github.com/jingxianna/ajps-dataset-analysis>, containing these two files along with python script and some other outputs (like file extension list).
- Two replication tales in Whole Tale.

THANK YOU!

Q & A