README for Replication Package

The Impact of Unions on Non-union Wage Setting: Threats and Bargaining

David A. Green[*] Ben M. Sand[†] Iain G. Snoddy[‡] Jeanne Tschopp[§]

October 2025

1 Paper Information

- **Title**: *The Impact of Unions on Non-union Wage Setting: Threats and Bargaining*
- **Authors**:
  - David A. Green — University of British Columbia and Research Fellow, IFS, London (`david.green@ubc.ca`)
  - Ben M. Sand — Department of Economics, York University (`bmsand@yorku.ca`)
  - Iain G. Snoddy — Analysis Group, Montreal (`iain.snoddy@analysisgroup.com`)
  - Jeanne Tschopp — University of Bern (`jeanne.tschopp@unibe.ch`)
- **Journal**: *Journal of Political Economy*
- **Corresponding Author**: David A. Green

2 Description of Files

File structure:

```
root/_Rep_Package/
├── Master.do
├── Master_for_R_Figures.R
└── Programs/
    ├── 01_cps
    ├── 01_May_matching
    ├── 01_IPUMS_CPS
    ├── 01_nlrb
    └── 01_cbp
```

[*]University of British Columbia and Research Fellow, IFS, London `david.green@ubc.ca`
[†]Department of Economics, York University `bmsand@yorku.ca`
[‡]Analysis Group, Montreal `iain.snoddy@analysisgroup.com`
[§]University of Bern `jeanne.tschopp@unibe.ch`

```
    ├── 02_preprocess
    ├── 03_analysis/
    │   ├── STATA_scripts
    │   └── Rscripts
├── raw_data/
    ├── CBP
    ├── cps
    ├── ipums_cps_matching
    ├── May_matching
    ├── partisanship
    └── NLRB
├── working_data/
    ├── CBP
    ├── cps
    ├── decomp_dynamic
    ├── keyvars
    ├── May_matching
    ├── ipums_ind_trans
    └── nlrb
├── temp
├── results/
    ├── tables
    └── figures
```

File descriptions:

File or Folder	Description
`Master.do`	Main script that replicates all tables and figures. Users can set global paths. Calls other scripts that loads & cleans data, processes variables, and generates all datasets required for tables. Also calls R scripts for final figure creation.
`Programs/01_cps/`	Scripts for extracting and cleaning CPS MORG data (1983–2019) from NBER and IPUMS.
`Programs/01_May_matching/`	Scripts for processing CPS May extracts (1977–1982) and matching with MORG samples.
`Programs/01_IPUMS_CPS/`	IPUMS CPS extraction for linking individuals in MORG.
`Programs/01_nlrb/NLRB Master.do`	Processes NLRB union election data provided by Henry Farber. Auxiliary files in `Programs/01_nlrb`

`Programs/01_cbp/Run_extract.do`	Processes County Business Patterns (CBP) data, including merging data from Eckert et al. [2021]
`Programs/02_preprocess/`	Scripts that construct key derived variables and prepare analysis datasets. Includes: `N_counts.do`, `ER.do`, `union_elections.do`, and `cps_transitions.do`.
`Programs/03_analysis/STATA_scripts/`	Stata scripts for all tables and decomposition figures. Includes: `Table_1.do` through `Table_8.do`, decomposition routines (`Decomposition_Data.do`, `Decomposition_Data_Chain.do`), and data generators for figures (`Figure_2_data.do`, etc.).
`Programs/03_analysis/Rscripts/`	R scripts for generating Figures 2–6 using data output from Stata. Run via `Master_for_R_Figures.R`.
`raw_data/`	Contains raw input data files.
`working_data/`	Stores cleaned intermediate datasets used across scripts.
`results/tables/`	Contains final .tex output for Tables 1–3 and appendix tables.
`results/figures/`	Contains .pdf output for Figures 2–6 and associated appendix figures.

3. Replication Instructions

Software Requirements

- STATA (tested on version 16). Ran on the following set up:

```
Stata/MP 16.1 for Mac (Apple Silicon)
Revision 13 Jun 2023
Copyright 1985-2019 StataCorp LLC
Total physical memory: 64.00 GB
Stata license: Single-user 8-core  perpetual
```

- R (version 4.3 or higher). Session information:[1]

```
version  R version 4.5.0 (2025-04-11)
os       macOS Sequoia 15.5
system   aarch64, darwin20
```

[1]The rest of the session information, including package versions, is included in the replication package as "R_session_info.txt".

```
 ui       RStudio
 date     2025-09-28
 rstudio  2025.05.0+496 Mariposa Orchid (desktop)
```

- R packages: `tidyverse`, `haven`, `latex2exp`, `zoo`, `ipumsr`, `AER`, `stringr`, `patchwork`, `DescTools`, `ggrepel`.

```
install.packages(c(
  "DescTools",
  "tidyverse",
  "haven",
  "latex2exp",
  "ggrepel",
  "zoo",
  "stringr",
  "AER",
  "patchwork"
))
```

- STATA packages: `reghdfe`, `estout`, `dm88_1`, `_gwtmean`
- `reghdfe` is used for estimation, see Correia [2016] for documentation and installation instructions.
- `estout` is used to construct latex tables, see Jann [2004]. We use a modified `.ado` file for the command `esttab`, and have included this custom file in the replication package.
- `dm88_1` is used for the command `renvars`.
- `_gwtmean` for `egen` style weighted means, see Kantor [2001].
- To install packages, this code can be used:

```
cap ado uninstall ftools
net install ftools, from("https://raw.githubusercontent.com/sergiocorreia/
ftools/master/src/")
cap ado uninstall ivreg2
ssc install ivreg2
cap ado uninstall ivreghdfe
net install ivreghdfe, from(https://raw.githubusercontent.com/sergiocorreia/
ivreghdfe/master/src/)
ssc install estout, replace
ssc install _gwtmean, replace
net install dm88_1, from(http://www.stata-journal.com/software/sj5-4)
```

To Replicate All Results

1. Download raw data from free, publicly available sources cited below in the Data Availability Statement and place in the corresponding `raw_data` folders.

2. Open `Master.do`. Edit `global root` to the full path of the replication folder on your own machine.

3. Open `Master_for_R_Figures.R`. Edit the `root` object to the full path of the replication folder on your own machine.

4. To run the R scripts directly from STATA, edit: `shell /usr/local/bin/Rscript` on line 270 of `Master.do` to the location of `Rscript` on your machine. Or just run `Master_for_R_Figures.R` in R after `Master.do` runs.

5. Outputs will be saved to `$root/results/tables` and `$root/results/figures`

6. Approximate runtime is 11 hours.

3 Data Availability Statement

This section describes how to access all raw data and reproduce the analytical datasets used in the paper. All data sources are publicly available, and no proprietary or confidential data are used.

- This replication package does not include raw microdata, which are subject to licensing and redistribution restrictions.

- All raw data used in the paper are publicly available at no cost from the providers listed below. To replicate, one must download these files directly and place them in the appropriate subfolders of `raw_data/`.

- All scripts required to clean and process these data are included. Derived datasets generated by these scripts are stored in `working_data/`.

Included vs. Excluded

- **Included**: All Stata and R scripts that are required to clean raw data, process, and construct tables and figures.

- **Excluded**: Raw CPS extracts (MORG, May), IPUMS CPS microdata, County Business Patterns (CBP), raw NLRB election case files, and Klarner datasets.

Data Sources

- **Current Population Survey, Merged Outgoing Rotation Group (MORG) from 1983-2019.** Retrieved August 2nd, 2017 and November 17th, 2020, from National Bureau of Economic Research (NBER). U.S. Census Bureau [2019].

  To replicate our analysis, download the annual CPS MORG files for 1983 through 2019 from the link provided. Each year's file is provided as a Stata-format dataset (e.g., `morg83.dta`, `morg84.dta`, ..., `morg19.dta`). After downloading, place the files into the folder:

  - `raw_data/cps/`

  The scripts in `Programs/01_cps/` assume that all MORG files for 1983–2019 are present in this directory.

- **Current Population Survey, May Extracts from 1977-1982**. Retrieved May 23rd, 2014, from National Bureau of Economic Research (NBER). U.S. Census Bureau [1982].

  To replicate our analysis, download the CPS May Extracts for survey years 1977 through 1982 from the NBER site. Each year's file is provided as a Stata dataset (e.g., `cpsmay77.dta`, `cpsmay78.dta`, ..., `cpsmay82.dta`). After downloading, place these files into the folder:

  - `raw_data/cps/`

  The scripts in `Programs/01_CPS/` assume that all May Extracts for 1977–1982 are present in this directory.

- **Matched Current Population Survey.** 1988-2020. Retrieved from IPUMS CPS on January 21st, 2025. Flood et al. [2024]. Due to IPUMS licensing restrictions, the raw microdata extract used in this paper is not redistributed in this replication package. Researchers must obtain the identical data directly from IPUMS by registering for a free account and recreating our extract. To do so, log into IPUMS CPS, select the same variables and years as in our project extract (provided in `Programs/01_IPUMS_CPS/`), and download both the data file (`.dat`) and the corresponding command file (`.do` or `.xml`) that IPUMS provides.

  After downloading, place the raw IPUMS extract file and its command file into the folder:

  - `raw_data/ipums_cps_matching`

  Our scripts expect the file names provided by IPUMS (e.g., `cps_00055.dat`). The Stata code in `Programs/01_IPUMS_CPS/` uses these files to load and process the extract into analysis-ready datasets.

- **Matched May CPS.** 1977-1982. Retrieved from IPUMS CPS on January 21st, 2025. Flood et al. [2024].

  Due to IPUMS licensing restrictions, the raw microdata extract used in this paper is not redistributed in this replication package. Researchers must obtain the identical data directly from IPUMS by registering for a free account and recreating our extract. To do so, log into IPUMS CPS, select the same variables and years as in our project extract (provided in `Programs/01_May_Mactching`), and download both the data file (`.dat`) and the corresponding command file (`.do` or `.xml`) that IPUMS provides.

  After downloading, place the raw IPUMS extract file and its command file into the folder:

  - `raw_data/may_matching/`

  Our scripts expect the file names provided by IPUMS (e.g., `cps_00058.dat`). The Stata code in `Programs/01_May_Mactching` uses these files to load and process the extract into analysis-ready datasets.

- **County Business Patterns (CBP), 1975–2010.** Retrieved from the U.S. Census Bureau and the U.S. National Archives (NARA). U.S. Census Bureau [1975–2010].

  To replicate our analysis, download the CBP data covering 1975–2010. The source and download procedure differ across time periods:

  - **1975–1985:** Files for these years are archived by the U.S. National Archives and Records Administration (NARA) and can be downloaded from their Amazon S3 storage system. Each year contains up to nine district-level text files named in the format: `RG029.CBPYY.T2I`*`D`*, where `YY` denotes the two-digit year and `D` is the district number (1–9). For example: `RG029.CBP74.T2I1`, `RG029.CBP80.T2I4`.

    A sample download URL (for 1980, District 1) is:
    https://s3.amazonaws.com/.../RG029.CBP80.T2I1.
  - **1986–2010:** Beginning in 1986, CBP data are hosted by the U.S. Census Bureau and can be downloaded directly from: https://www2.census.gov/programs-surveys/cbp/data/datasets/. Each annual county-level file is provided as a compressed ZIP archive named: `cbpYYco.zip`, where `YY` is the two-digit year (e.g., `cbp86co.zip`, `cbp99co.zip`, `cbp00co.zip`, ..., `cbp10co.zip`).

    After downloading, unzip each archive into a subfolder named after the file (e.g., `cbp86co/`).

  After downloading (and unzipping for post-1985 data), place all CBP files into:

  - `.../raw_data/cbp/`

- Imputed CBP missing values from Fabian Eckert, Teresa C. Fort, Peter K. Schott, and Natalie J. Yang (2021). Details and documentation contained in Eckert et al. [2021].

  - `.../raw_data/cbp/efsy/X/Final Imputed/efsy_cbp_X.csv` for `X` in 75, ..., 10.

We also use their industry concordances, which must be placed in the folder:

  - `.../raw_data/cbp/`

These imputed CBP files are publicly available and should be downloaded directly from the authors' website; we do not redistribute them here.

- Our NLRB data comes from two sources:

  1. NLRB 1977-1999. Data and documentation can be found on Thomas J. Holmes's data page. Part of original source comes from Henry Farber. The construction of the data is described in Holmes [2006a,b].

     - `.../raw_data/nlrb/nlrb7799.dta`

  2. NLRB 1999–2010. Data and documentation are available on John-Paul Ferguson's GitHub repository [Ferguson, 2013]. This repository hosts copies of publicly available data from the National Labor Relations Board (NLRB). The underlying source data originate from the NLRB's public case-tracking databases (CATS and associated systems), and are provided as zipped XML files.

     Replication Steps. The following steps describe how to reproduce the NLRB datasets used in our analysis. Steps (1)–(5) must be completed *before* running `Master.do`.

     - Download the annual XML archives (1999–2010) from John-Paul Ferguson's GitHub repository. Filenames include `R_YYYY.zip`, `ATS_R_CASE_Data_YYYY.zip`, or `CATS-FRF-R-YYYY.final.zip`.[2] Save all files to `raw_data/nlrb/`.
     - Unzip each archive in the same directory: `raw_data/nlrb/`.
     - Convert the XML files to CSV format using Ferguson's Bash script `clean_R.sh` and instructions. This results in CSV files being placed in `raw_data/nlrb/`.
     - Ferguson also provides a Stata cleaning script, `CleanNLRBcsvs.do`. We adapt this code and include our own Stata script to clean the CSV files. Our script depends on several `.ado` files supplied by Ferguson, which should be downloaded and placed in `programs/01_nlrb/`. These are automatically loaded when `CSV_to_DTA.do` is called from `Master.do`.
     - Download the file `nlrb_representation_1961_1998.dta.zip` from Ferguson's GitHub and unzip to `raw_data/nlrb/`.
     - The above steps must be taken prior to running our `Master.do` file, which calls on `programs/01_nlrb/NLRB Master.do`, which converts the raw CSV files into Stata-format data that correspond to:
       * `CATS-R-FRF.dta` (case-level information)
       * `R_CLOSED_CASE.dta` (case outcomes and closure reasons)
       * `R_ELECT_CERTIFICATION.dta` (certification decisions)
       * `R_ELECT_VOTES_FOR.dta` (vote totals by outcome)

[2] e.g., `https://github.com/jpfergongithub/nlrb-cats/raw/master/R_2001.zip`

    * `R_ELECTION_TALLY.dta` (summary of election results)

  and are placed in the folder `raw_data/nlrb` and used in our analysis.

3. *State Partisan Balance Data, 1937–2011*, from Carl Klarner Dataverse. Klarner [2013]
   - Download the file `Partisan_Balance_For_Use2011_06_09b.xlsx`
   - Place file in `raw_data/partisanship`
4. *State Partisan Composition* from Phillip Thomas. Thomas [2024]. The data are organized by state and year, and have counts for the total, Democratic, Republican, and other members for each legislative body. Original source from Klarner [2013].
   - Download the file `state_partisan_composition_1934_2021.csv`
   - Place file in `raw_data/partisanship`

4 Known Issues or Warnings

- We've used a modified `esttab.ado` file, provided in the package and loaded in the `Master.do` file.

5 Citation

If you use this code or data, please cite:

> Green, Sand, Snoddy, and Tschopp. "*The Impact of Unions on Non-union Wage Setting: Threats and Bargaining.*" *Journal of Political Economy*, forthcoming.

6 Licensing

All data used in this project originate from publicly available sources listed in the Data Availability Statement. The authors have the right to use all data referenced in this package. The replication package includes only the original code that reproduces the full underlying microdata derived from raw sources.

Data Licensing and Redistribution

- **Current Population Survey (CPS) data**: Produced by the U.S. Census Bureau and Bureau of Labor Statistics. Distributed by the National Bureau of Economic Research (NBER) and IPUMS CPS. Redistribution of the raw microdata is not permitted under NBER and IPUMS licensing terms. Researchers must obtain the CPS data directly from those sources using the links and instructions provided in the Data Availability Statement.

- **County Business Patterns (CBP)**: Produced and distributed by the U.S. Census Bureau. Publicly available for download at no cost. Raw CBP data are excluded from this package and must be obtained from the Census Bureau website.
- **NLRB election data**: Portions of these data were obtained from Thomas J. Holmes's personal webpage (University of Minnesota) and from John-Paul Ferguson's GitHub repository. Both sources host publicly available copies of data originally released by the National Labor Relations Board (NLRB). These data are cited and linked in the Data Availability Statement but not redistributed here.
- **IPUMS CPS extracts**: Distributed by IPUMS under its user agreement. Because redistribution of complete extracts is restricted, only variable definitions and extract specifications are provided. Replicators must create their own IPUMS extract using the provided specifications.
- **Klarner Partisan Balance Data**: Distributed via the Carl Klarner Dataverse under a public-use license. Users should cite the dataset as indicated in the Dataverse record.
- **Phillip Thomas Partisan Composition Data**: From Phillip Thomas's GitHub repository. These data are cited and linked in the Data Availability Statement but not redistributed here.

Code

All Stata and R scripts in this package are released under a permissive license (CC BY 4.0 or equivalent). Users may copy and modify the code with attribution to the original authors.

References

S. Correia. Linear models with high-dimensional fixed effects: An efficient and feasible estimator. Technical report, 2016. Working Paper.

F. Eckert, T. C. Fort, P. K. Schott, and N. J. Yang. Imputing missing values in the us census bureau's county business patterns. Technical report, National Bureau of Economic Research, 2021.

J.-P. Ferguson. Nlrb case data repository, 2013. [dataset] GitHub repository. Available at https://github.com/jpfergongithub. Accessed October 15, 2025.

S. Flood, M. King, R. Rodgers, S. Ruggles, J. R. Warren, D. Backman, A. Chen, G. Cooper, S. Richards, M. Schouweiler, and M. Westberry. Ipums cps: Version 12.0, 2024. URL https://doi.org/10.18128/D030.V12.0. [dataset].

T. J. Holmes. Geographic spillover of unionism. *NBER Working Paper Series*, 2006a. URL https://api.semanticscholar.org/CorpusID:3105553.

T. J. Holmes. Union election data, 1977–1999, 2006b. [dataset] Personal webpage, University of Minnesota. Accessed October 15, 2025. Available at https://users.econ.umn.edu/~holmes/data/geo_spill/.

B. Jann. Estout: Stata module to make regression tables. Statistical Software Components, Boston College Department of Economics, Jul 2004. URL https://ideas.repec.org/c/boc/bocode/s439301.html.

D. Kantor. _gwtmean: Stata module containing extensions to generate to implement weighted mean. Statistical Software Components, Boston College Department of Economics, Jun 2001. URL https://ideas.repec.org/c/boc/bocode/s418804.html.

C. Klarner. State Partisan Balance Data, 1937-2011, 2013. URL https://doi.org/10.7910/DVN/LZHMG3. Accessed: October 15th 2025.

P. Thomas. State partisan composition data. https://github.com/psthomas/state-partisan-composition, 2024. Accessed: October 15th 2025.

U.S. Census Bureau. County business patterns (cbp), 1975–2010. URL https://www.census.gov/programs-surveys/cbp.html. [dataset] Retrieved October 15th, 2025.

U.S. Census Bureau. Current population survey: May extracts, 1977–1982, 1982. URL https://www.nber.org/research/data/current-population-survey-cps-basic-monthly-data. [dataset] Distributed by the National Bureau of Economic Research (NBER). Accessed October 15, 2025. Last updated August 2025. DOI: https://doi.org/10.60592/d7gd-cx91.

U.S. Census Bureau. Current population survey: Merged outgoing rotation group (morg), 1983–2019, 2019. URL https://www.nber.org/research/data/current-population-survey-cps-basic-monthly-data. [dataset] Distributed by the National Bureau of Economic Research (NBER). Accessed October 15th, 2025. Last updated August 2025. DOI: https://doi.org/10.60592/d7gd-cx91.