

Accounting Rules and the Labor Market for Accountants

Data Description Sheet and ReadMe File

1 Data Description Sheet

1. **A description of which author(s) handled the data and conducted the analyses.**

Anthony Le handled the data and conducted the analyses.

2. **A detailed description of how the raw data were obtained or generated, including data sources, the specific date(s) on which data were downloaded or obtained, and the instrument used to generate the data (e.g., for surveys or experiments). We recommend that more than one author is able to vouch for the stated source of the raw data.**

A listing of the data sources is provided below:

- (a) **Occupational Wage and Employment Statistics (OEWS):** These data are publicly available from the [following database](#). These data were downloaded in June of 2023.
- (b) **Revelio Labs:** These data were obtained via the WRDS database at the University of Chicago. These data were downloaded in July of 2025.
- (c) **Online Job Postings Data:** These data were obtained from a proprietary data provide, Linkup, through an agreement with Columbia Business School's Accounting Division. These data were downloaded in August of 2023.
- (d) **FASB Standards Data:** Versions of the FASB standards were obtained from the FASB's Reference Library, which can be [found here](#). The data on each standard's share of restrictive words was obtained from Hribar et al. (2022). These data were downloaded in November of 2022.
- (e) **Compustat Data:** These data were obtained through Wharton Research Data Services (WRDS) through a proprietary agreement with Columbia University. These data were downloaded in July of 2023.
- (f) **Survey Data and Survey-Based Field Experiment:** These data were distributed and collected via an online survey run on Qualtrics. The full survey instruments are shown in Online Appendix Section A1. The survey data collection took place in September of 2023.
- (g) **Glassdoor Data:** These data were obtained from Glassdoor in July of 2024.

(h) **National Survey of College Graduates:** These data were obtained from the [following database](#) in August of 2023.

3. If the data are obtained from an organization on a proprietary basis, the authors should privately provide the editors with contact information for a representative of the organization who can confirm data were obtained by the authors. The editors would not make this information publicly available. The authors should also provide information to the editors about the data sharing agreement with the organization (e.g., non-disclosure agreements, any restrictions imposed by the organization on the authors, such as restrictions to publish certain results). In particular, the authors should indicate if an organization or data provider imposes restrictions on the publication of the results, has not given the authors full control of the relevant data, requires that the results have to be reviewed or approved prior to public release of the paper or publication. This information should be provided to the editors upon submission.

The data used in this draft of the paper is all publicly available or are commercially licensed data.

4. A complete description of the steps necessary to download, obtain or collect as well as process the data used in the final analyses reported in the paper. For experimental and survey papers, we require information about the instructions and instruments used to generate the data, subject eligibility and/or selection, as well as any exclusion criteria. The full set of instructions and instruments can be provided in the online appendix.

See the README file below, the accompanying code, and the online appendix for relevant documentation.

5. After downloading or obtaining the raw data, all manipulations of the data should be done via computer programs. The code for these manipulations should be included in the code submitted upon acceptance (see below). No manipulations of raw data can take place manually or outside the computer code provided. If compliance with this requirement is not feasible, the authors need to explain and disclose any manipulations of the raw data (e.g., manually created variables or file conversions). When feasible, we also encourage the authors to share the code that downloads the data.

See the README file below, the accompanying code, and the online appendix for relevant documentation.

6. The computer programs (i.e., code) used to (1) convert the raw data into the final dataset used in the analysis, (2) to execute the statistical or econometric analysis, and (3) to generate the tables or to produce the output used in constructing tables of the manuscript. A brief description that enables other researchers to understand and run the code should be provided. The purpose of this requirement is to facilitate replication and to help other researchers understand in detail how the raw data were processed, the final sample was formed, variables were defined, outliers

were treated, and which commands were used in the analysis, etc. This code or programming is in most circumstances not proprietary. However, we recognize that some parts of the code or data generation process may be proprietary, including from the authors' perspective. Therefore, instead of disclosing the proprietary portion of the code or program, researchers can provide a detailed step-by-step description of the code or the relevant parts of the code such that it enables other researchers to arrive at the same results that the authors obtained and presented in their manuscript. In such cases, the authors should inform the editors upon initial submission, so that the editors can consider an exemption allowing the step-by-step description. Whenever feasible, authors are required to provide the identifiers (e.g., CIK, CUSIP) for their final sample. Authors should consult our FAQ Sheet on the JAR website for further details.

See the README file below, the accompanying code, and the online appendix for relevant documentation.

7. **A comprehensive log file that shows the execution of the entire code.** This log file should cover all the steps that convert the raw data into a final dataset and the execution of all statistical and econometric analyses presented in the tables of the manuscript. The portion of the log file that shows proprietary code or data may be masked. In this case, the reader should be referred to the step-by-step description provided as per the requirements in Item 6.

See the README file below, the accompanying code, and the online appendix for relevant documentation.

8. **An assurance that the data and programs will be maintained by at least one author (usually the corresponding author) for at least six years, consistent with National Science Foundation guidelines.**

The data and programs will be maintained by the corresponding author for at least six years, consistent with the NSF guidelines.

2 Overview and Data Structure of Replication Package

- This file describes the structure of the directory to replicate the results from **Accounting Rules and the Labor Market for Accountants**
- The replication package contains three main folders:
 - **01 - Code:** Contains all of the code to transform the data from its raw form through to the analysis.
 - **02 - Data:** Contains all of the raw, intermediate, and output data files.¹
 - **03 - Results:** Stores log files outputted from the code and directories where tables/figures would be outputted.

3 Description of Folders in Replication Package

Folder 1: 01 - Code

The “01 - Code” folder is organized such that the code files for different sets of analyses (and datasets) are contained within their respective folders. To reproduce the results from the main text in their entirety, you can use master.do and the notes contained within the master do-file (which requires running a couple of separate programs using Python). Below is a brief description of the files and subfolders in the “01 - Code” folder, roughly in the order they appear in the master do file:

- **master.do** — This is the master do-file. If attempting to reproduce all the results from the paper, please run this do-file only and it will contain all code necessary to reproduce all Tables/Figures in the main text. To run on your machine, please make sure to change pathnames in the master.do file.
- **Treatment Measure and Descriptives** — This folder contains all of the code needed to construct the main treatment measures used in the manuscript (*Restrictiveness*, *Exposure to Restrictive Words*, *Exposure to Total Words*). The folder contains the following code files:
 - **firm_standard_exposure.ipynb:** This file computes the exposure of firms to each of the FASB standards using the dictionary in Folsom et al. (2017).
 - **state_level_restrictiveness_measure.do:** This file creates the treatment measures used in the manuscript. The file also reproduces Panels A and B of Figure 2.
 - **maps.do:** This file generates the descriptive maps shown in the manuscript, including Figure 1 and Figure 3.
 - **industry_descriptive_statistics.do:** This file generates the sector-level descriptive statistics shown in Panel B of Table 1.

¹I cannot include data in this replication package because most of the data can only be accessed via agreements with the data vendor. Nevertheless, I have structured the directories in the replication to match the directory structure required by the code. Below, I also provide detailed descriptions of the content of each raw, intermediate, and output data file that would be in each of these directories.

- **OEWS** — This folder contains the code needed to reproduce the results using the Occupational Employment and Wage Statistics (OEWS) database. The folder contains the following code files:
 - **OEWS_employment.do**: This file combines the individual annual-level OEWS files and reproduces the results from Panel A of Table 2.
 - **OEWS_wages.do**: This file reproduces the results using wage data from the OEWS in Table 5.
- **HERI** — This folder contains the code needed to reproduce the results using the Higher Education Research Institute’s (HERI) The Freshman Survey (TFS) database. The folder contains the following code files:
 - **student_major_choice**: This file creates the analysis dataset of students from the HERI data and reproduces the results in Panel B of Table 2.
 - **outreach_programs_cross_section.do**: This file reproduces the results using outreach programs as a cross-sectional test in Panel A of Table 6.
 - **accounting_parent_cross_section.do**: This file reproduces the results using accountant parents as a cross-sectional test in Panel B of Table 6.
- **Revelio** — This folder contains the code needed to reproduce the results using the Revelio Labs database. The folder contains the following code files:
 - **revelio_job_switches**: This file creates the sample from the Revelio Labs data and reproduces the results from Panels A, B, and C of Table 3.
- **LinkUp** — This folder contains the code needed to reproduce the results using the LinkUp database. The folder contains the following code files:
 - **linkup_header.do**: This file creates a header file of unique hashes from the raw job records database from LinkUp.
 - **XML_Convert_job_descriptions.ipynb**: This file converts the XML job descriptions file into a tabular format, identifying job descriptions via their unique hashes from the header file created by linkup_header.do.
 - **job_description_word_counts.ipynb**: This file conducts a word count of terms from the Atalay et al. dictionary for each job description in the LinkUp database.
 - **analysis_labor_demand.do**: This file combines and appends the individual word count data created by the previous code file, and conducts the analysis that reproduces Panels A and B of Table 4.
- **Other** — This folder contains other miscellaneous archival tests in the manuscript that are not already covered by the other files. This includes, for example, the Glassdoor analysis and analysis using the National Survey of College Graduates (NSCG). The folder contains the following code files:
 - **glassdoor_reviews.do**: This code file creates the Glassdoor sample and reproduces the results from Panel A of Table 7.

- **nscg.do**: This code file creates a regional version of my treatment measure (since the NSCG data is only at the region level and not the state level), and reproduces the results in Panel B of Table 7.
- **Survey and Experiment** — This folder contains other the code needed to reproduce the results of the survey and survey-based experiment:
 - **survey.do**: This code file reproduces the descriptive statistics for the survey tabulated in Table 8, and the descriptive figures shown in Figure 4.
 - **experiment.do**: This code file reproduces the results from the survey-based experiment, shown in Table 9.

Folder 2: “02 - Data”

Similar to the “01 - Code” folder, the “02 - Data” folder is organized such that the different datasets are contained within their respective folders. In each of these folders, there are three subfolders: orig (raw data), intermediate (intermediate datasets used across code files), and output (final datasets used for analysis across code files). Below is a brief description of the files and subfolders in the “02 - Data” folder:

- **Occupational Employment & Wage Statistics (OEWS):**
 - orig:
 1. **state_XXXX_dl.xls** (where XXXX is the year between 1997 - 2019):
These data are the raw datasets collected from the [OEWS database](#).
 - intermediate:
 - output
 1. **OEWS_analysis.dta**: This dataset is created by the file OEWS_employment.do and contains all the data needed to run the analysis using the OEWS data.
- **Higher Education Research Institute (HERI):**
 - orig:
 1. **TFS/DISAGG.dta**: This dataset contains data from 1961 - 2009, obtained via [HERI's](#) The Freshman Survey (TFS) database. This data contains basic ID variables and information on each student covered by the TFS.
 2. **TFS/DEMOGRAPHICS.dta**: This dataset contains data from 1961 - 2009, obtained via [HERI's](#) The Freshman Survey (TFS) database. This data contains basic demographic information for each student covered in the TFS database.
 - intermediate:
 - output:
 1. **merged_student_data.dta**: This dataset is created by the file student_major_choice.do. It contains the cleaned student sample from the DISAGG data, merged with the DEMOGRAPHIC data, before it is collapsed to the Major-State-Year level.
- **Revelio Labs:**

- orig:
 1. **individual_positions.csv**: Raw positions dataset obtained from WRDS Revelio Labs database.
 2. **lookup.csv**: Bridge between ONET codes and Revelio’s k1500 occupation codes.
- intermediate:
- output:
 1. **switches_weighted.dta**: This dataset contains the number of job switches by an employee’s current occupation, grouped into three categories: non-accounting business occupations, accounting positions, and other non-business occupations.
 - * Created by 01 - Code/Revelio/revelio_job_switches.do.
 2. **job_switches_seniority_weighted.dta**: This dataset contains the number of job switches into any new occupation, broken down by seniority and current occupation.
 - * Created by 01 - Code/Revelio/revelio_job_switches.do.
 3. **job_switches_seniority_weighted_othbiz.dta**: This dataset contains the number of job switches into non-accounting business occupations broken down by seniority and current occupation.
 - * Created by 01 - Code/Revelio/revelio_job_switches.do.

• LinkUp:

- orig:
 1. **raw_job_archive_2020-12-31/raw_job_archive_2020-12-31_XXXX.csv**: These are individual job records data obtained from LinkUp, which contains all job postings in the LinkUp database up to, and including, December 31st, 2020.
 2. **raw_description_data/linkup_job_descriptions_2020-12-31_XXXX**: These are individual job description data obtained from LinkUp, which contains descriptions for all job postings in the LinkUp database up to, and including, December 31st, 2020.
 - * **NOTE**: Place XML_Convert_job_descriptions.ipynb in the orig folder to run to output job descriptions database.
- intermediate:
 1. **input/output_XXXX.csv**: Contains the outputted file from 01 - Code/XML_Convert_job_descriptions.ipynb, chunked into 76 separate files, now to be used as an input for the word counts of Atalay et al. (2020) dictionary.
 - * **NOTE**: Place job_description_word_counts.ipynb in the intermediate folder to run to output word counts from Atalay et al. (2020) dictionary. Also place dictionary spreadsheet atalaywords.xlsx in this directory.
 2. **output/output_XXXX.csv**: Outputted word counts from Atalay et al. (2020) dictionary, produced by 01 - Code/XML_Convert_job_descriptions.ipynb, chunked into 76 separate files.
- output:

1. **job_postings_data_combined.dta**: Contains the appended raw job records files.
2. **job_posting_descriptions_individual_amenities.dta**: Contains the appended 76 Atalay et. al (2020) word counts files.
3. **raw_job_postings_and_descriptions_individual_amenities_ind.dta**: Contains analysis sample for job postings analysis.

- **Other:**

- orig:

1. **10KQ/YEAR/XXXX.txt**: Each subfolder contains all 10-Ks and 10-Qs filed in the “YEAR” of the subfolder.
 - * **Note:** Place `firm_standard_exposure.ipynb` in this folder to run to create firm-level standard exposures.
2. **CBP/bcp2000st.txt**: Obtained from the County Business Patterns database and contains data on the industrial composition of states for 2000.
3. **CBP/efsy_naics97_to_naics12.csv**: Crosswalk from NAICS 1997 codes to NAICS 2012 codes.
4. **CBP/state_names.txt**: Crosswalk of FIPS State to State Acronyms
5. **CBP/state_to_region.xlsx**: Crosswalk of state to NSCG regions.
6. **Compustat/compustat_fund_1990_2020.dta**: Compustat fundamentals downloaded from WRDS for all available industrial firms from 1990 to 2020.
7. **FASB Standards/share_keywords.xlsx**: Standard-level share of restrictive words measure from Hribar et al. (2022).
8. **FASB Standards/std_year.sas7bdat**: Standard-level count of restrictive words measure from Hribar et al. (2022).
9. **map_data/s_08mr23/**: United States shapefile for maps
10. **map_data/state_coordinates.dta**: United States coordinate data for map
11. **map_data/state_db.dta**: United States master database for maps
12. **NSCG/ecg03.sas7bdat**: NSCG 2003 wave
13. **NSCG/ecg10.sas7bdat**: NSCG 2010 wave
14. **NSCG/ecg13.sas7bdat**: NSCG 2013 wave
15. **NSCG/ecg15.sas7bdat**: NSCG 2015 wave
16. **NSCG/ecg17.sas7bdat**: NSCG 2017 wave
17. **outreach_programs_updated.xlsx**: Data on accounting outreach programs and their introduction across different states, hand-collected by visiting states’ archived websites via the WayBack machine.
18. **reviews_external_unc.csv**: A raw dataset of all Glassdoor reviews

- intermediate:

1. **firm_standard_keyword_counts/firm_standard_keyword_counts_XXXX.csv**: Contains standard-firm-year counts of keywords from Folsom et al. (2017) for all 10-Ks within a year.

- * Created by 01 - Code/Treatment Measure and Descriptives/
firm_standard_exposure.ipynb
- output:
 1. **appended_10K_rel_imp.dta**: Appended version of relative importance of each standard to each firm-year observations across all years.
 - * Created by 01-Code/Treatment Measure and Descriptives/
state_level_restrictiveness_measure.do
 2. **industry_year_restrictiveness.dta**: Average industry-year exposure to restrictiveness across all firms within a 4-digit NAICS industry.
 - * Created by 01-Code/Treatment Measure and Descriptives/
state_level_restrictiveness_measure.do
 3. **region_restrictiveness_data.dta**: Contains region-year level exposure to GAAP restrictiveness.
 - * Created by 01-Code/Other/nscg.do
 4. **restrictiveness_data.dta**: Contains state-year level exposure to GAAP restrictiveness (i.e., main treatment measure).
 - * Created by 01-Code/Treatment Measure and Descriptives/
state_level_restrictiveness_measure.do
- **IDs**: This folder contains the IDs that are included in the final sample. Since the analysis is done on the aggregate level (often the occupation-state-year level), identifiers are shown for the relevant aggregation level for the analysis dataset.

Folder 3: “03 - Results”

The “03 - Results” folder would contain all output generated by the code files in “01 - Code.” The folder is organized into Figures, Tables, and Log files (which are created for each do-file, but can also be generated all at once using the master.do file).² Below is a brief description of the subfolders, and the files that should be created within those subfolders via the code, in the “03 - Results” folder:

- **Figures** — The figures from the main text that are outputted by the code base should consist of the following:
 1. **treatment_map_1995.png** (Panel A of Figure 1)
 - Created by 01-Code/Treatment Measure and Descriptives/maps.do
 2. **treatment_map_2005.png** (Panel B of Figure 1)
 - Created by 01-Code/Treatment Measure and Descriptives/maps.do
 3. **treatment_map_2015.png** (Panel C of Figure 1)
 - Created by 01-Code/Treatment Measure and Descriptives/maps.do
 4. **time_series_variation.pdf** (Panel A of Figure 2)
 - Created by 01-Code/Treatment Measure and Descriptives/
state_level_restrictiveness_measure.do

²For the purposes of this replication package, only the log files are produced as output, but all Figures/Tables in the main text can be produced by running the code base.

5. **cross_sectional_treatment_variation.pdf** (Panel B of Figure 2)
 - Created by 01-Code/Treatment Measure and Descriptives/
state_level_restrictiveness_measure.do
 6. **accountant_map_2000.png** (Panel A of Figure 3)
 - Created by 01-Code/Treatment Measure and Descriptives/maps.do
 7. **AcctStudents_map_2000.pdf** (Panel B of Figure 3)
 - Created by 01-Code/Treatment Measure and Descriptives/maps.do
 8. **deterrents.pdf** (Panel A of Figure 4)
 - Created by 01-Code/Survey and Experiment/survey.do
 9. **recommendation.pdf** (Panel B of Figure 4)
 - Created by 01-Code/Survey and Experiment/survey.do
 10. **accountant_demand.pdf** (Panel C of Figure 4)
 - Created by 01-Code/Survey and Experiment/survey.do
 11. **accountant_pay.pdf** (Panel D of Figure 4)
 - Created by 01-Code/Survey and Experiment/survey.do
 12. **source_information.pdf** (Panel E of Figure 4)
 - Created by 01-Code/Survey and Experiment/survey.do
- **Logs** — The log-files produced from each do-file. Full log files of each code file in this replication package is provided in this folder, including:
 1. accounting_parent_cross_section.log
 2. analysis_labor_demand.log
 3. experiment.log
 4. glassdoor_reviews.log
 5. maps.log
 6. nscg.log
 7. OEWS_employment.log
 8. OEWS_wages.log
 9. outreach_programs_cross_section.log
 10. revelio_job_switches.log
 11. state_level_restrictiveness_measure.log
 12. student_major_choice.log
 13. survey.log
 - **Tables** — The tables from the main text that are outputted by the code should consist of the following³:

³The only table not outputted below is Table 1, which requires special formatting. The descriptive statistics shown in Panel A Table 1 are shown at the end of each log file, and 01-Code/Treatment Measure and Descriptive/industry_descriptive_statistics.do creates Panel B of Table 1.

1. **Table 2A_Empl.tex** (Panel A of Table 2)
 - Created by 01-Code/OEWS/OEWS_employment.do
2. **Table2B_Majors.tex** (Panel B of Table 2)
 - Created by 01-Code/HERI/student_major_choice.do
3. **Table3A_Revelio.tex** (Panel A of Table 3)
 - Created by 01-Code/Revelio/revelio_job_switches.do
4. **Table3B_Revelio.tex** (Panel B of Table 3)
 - Created by 01-Code/Revelio/revelio_job_switches.do
5. **Table3C_Revelio.tex** (Panel C of Table 3)
 - Created by 01-Code/Revelio/revelio_job_switches.do
6. **Table4A_LinkUp.tex** (Panel A of Table 4)
 - Created by 01-Code/LinkUp/analysis_labor_demand.do
7. **Table4B_LinkUp.tex** (Panel B of Table 4)
 - Created by 01-Code/LinkUp/analysis_labor_demand.do
8. **Table5A_annualsalaries.tex** (Panel A of Table 5)
 - Created by 01-Code/OEWS/OEWS_wages.do
9. **Table5B_wage_dispersion.tex** (Panel B of Table 5)
 - Created by 01-Code/OEWS/OEWS_wages.do
10. **Table6A_Outreach.tex** (Panel A of Table 6)
 - Created by 01-Code/HERI/outreach_programs_cross_section.do
11. **Table6B_Parents.tex** (Panel B of Table 6)
 - Created by 01-Code/HERI/accounting_parent_cross_section.do
12. **Table7A_Glassdoor.tex** (Panel A of Table 7)
 - Created by 01-Code/Other/glassdoor_reviews.do
13. **Table7B_NSCG_satisfaction.tex** (Panel B of Table 7)
 - Created by 01-Code/Other/nscg.do
14. **survey_sumstats.tex** (Table 8)
 - Created by 01-Code/Survey and Experiment/survey.do
15. **Table9A_creativity.tex** (Panel A of Table 9)
 - Created by 01-Code/Survey and Experiment/experiment.do
16. **Table9B_noncreativity.tex** (Panel B of Table 9)
 - Created by 01-Code/Survey and Experiment/experiment.do

4 References

Atalay, Enghin, Phai Phongthientham, Sebastian Sotelo, and Daniel Tannenbaum, 2020, The evolution of work in the united states, *American Economic Journal: Applied Economics* 12, 1–34.

Folsom, David, Paul Hribar, Richard Mergenthaler, and Kyle Peterson, 2017, Principles-based standards and earnings attributes, *Management Science* 63, 2397–2771.

Hribar, Paul, Richard Mergenthaler, Aaron Roeschley, Spencer Young, and Chris X. Zhao, 2022, Do managers issue more voluntary disclosure when gaap limits their reporting discretion in financial statements?, *Journal of Accounting Research* 60, 299–351.