

Data Description Sheet

Diversity Tokenism

Kelvin Law and Jingdan Tan
August 2025

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

Kelvin Law and Jingdan Tan jointly handled the data and conducted the analyses in the paper.

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 can vouch for the stated *source* of the raw data.

The raw data for this study were obtained from external sources or generated by the authors as follows:

A. Externally Sourced Data:

- **BLM Protest Data:** Sourced from Dunivin, Yan, Ince, and Rojas (2022), which compiles data from Elephrame and the Armed Conflict Location & Event Data Project (ACLED). Obtained in November 2022.
- **Firm Headquarters:** Historical headquarters information from the WRDS SEC Analytic Suite. Obtained in January 2022.
- **Director and Executive Names and Records:** Sourced from ExecuComp. Obtained in February 2023.
- **Employee Information:** Workforce data from Revelio Labs. Obtained in September 2022.
- **Demographic Data:** County and state-level demographics from the American Community Survey. Obtained in November 2023.
- **Firm Fundamentals:** Sourced from Compustat. Obtained in August 2022.
- **Diversity Ratings:** Diversity and Inclusion Ratings from Refinitiv; Obtained in December 2022.
- **Reputational Risk:** Incident data from RepRisk. Obtained in November 2022.
- **Election Data:** Sourced from the MIT Election Lab. Obtained in January 2023.
- **Economic Indicators:** State Coincident Index data from the Federal Reserve Bank of Philadelphia (obtained December 2023) and employment data from the Bureau of Labor Statistics (obtained April 2023).

- **Employee-related Ratings:** Employee-related ESG scores (e.g., turnover, well-being) from S&P Global; Obtained in June 2025.
- **10-Ks:** Full text of raw 10-Ks. Obtained from The Notre Dame Software Repository for Accounting and Finance (SRAF).

B. Author-Generated Data:

- **Race/Ethnicity Classification Data:** We generated race/ethnicity classifications for directors and executives by using their names as inputs for OpenAI’s GPT-4 model via the ChatGPT web interface between March and April 2023. The process, which involved Chain-of-Thought prompting, is detailed in Section 3.2 and Online Appendix A.1.

Both authors can vouch for the source of the raw data.

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, and any restrictions imposed by the organization on the authors). 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 must be reviewed or approved prior to public release of the paper or publication.**

BLM protest data, county- and state-level demographics, economic indicators, election data, and 10-K filings are in the public domain and publicly accessible.

All other datasets, including Compustat, ExecuComp, Refinitiv, RepRisk, S&P Global, were accessed via institutional subscriptions through Nanyang Technological University during the course of the project.

Revelio Labs granted us a non-transferable, personal license to access and analyze their proprietary data. Redistribution or transfer of these data to any third party is prohibited under the licensing agreement.

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.**

The primary data processing steps were as follows:

- 1) **Merging:** We merged the various datasets at the firm-year level using gvkey and CIK identifiers.
- 2) **Race/Ethnicity Classification:** We developed and validated a novel classification approach using OpenAI’s GPT-4 to infer the race/ethnicity of directors and executives from their names, as detailed in Section 3.2. For workforce data, we utilized the classifications provided by Revelio Labs.
- 3) **Variable Construction:** We constructed our main dependent and control variables. This included calculating diversity percentages, Herfindahl-Hirschman Indices (HHI) for diversity, and our “Alignment Scores” which benchmark internal firm diversity against community demographics.
- 4) **Sample Construction and Weighting:** We constructed our main staggered difference-in-differences sample and our stacked DiD sample for robustness tests. For our primary analyses, we generated and applied entropy balancing weights to ensure covariate balance between treated and control groups based on pre-treatment county-level demographic characteristics.

Further details on these procedures are available in Section 3 and the Online Appendix of the manuscript.

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.**

We used Stata to process raw data and conduct statistical analyses. All Stata data manipulations are provided in the file *LT-sample-construction.do*.

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.

The Stata code used to convert raw data into the final datasets is provided in *LT-sample-construction.do*.

The Stata code used to execute statistical analyses and generate manuscript tables is provided in *LT-replacement probabilities.do*, and *LT-regressions.do*. Stata and Python code used to generate figures is provided in *LT-protest maps.do*, *LT-DiD plots.do*, and *LT-eco mag plot.py*.

We also provide a list of firm identifiers (*LT-identifier.dta*) in gvkey form for our final sample.

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.**

We provide comprehensive Stata log files that show the execution of the codes.

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.**

Both authors assure that the data and programs will be maintained for a minimum of six years.