

Data Description Sheet

“Human Capital Disclosure and Labor Market Outcomes: Evidence from Regulation S-K”

By Jung Ho Choi, Dan Li, and Daniele Macciocchi

This data description sheet is provided to fulfill the requirements of the Journal of Accounting Research data policy for submission of a manuscript.

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

Jung Ho Choi, Hannah Han, and Khanh Vu handled the ChatGPT data and process (Hannah Han and Khanh Vu are research assistants under Jungho Choi’s guidance). The rest of the data was handled by Dan Li, who also processed the entire data and conducted the analyses.

Item 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 datasets that we used in this paper are CRSP Mutual Fund Holdings Database, Morningstar’s annual Sustainable Funds U.S. Landscape Report, Factset, Revelio Labs, Capital IQ ESG Score, Capital IQ Transcripts, Compustat, RavenPack, LinkUp, and GlassDoor.

A detailed description of how the raw data were obtained and generated is presented in Section 3 and Appendix A of the manuscript.

We downloaded Compustat from Wharton Research Data Services (WRDS) in October 2022. We downloaded RavenPack in December 2022. We downloaded Revelio Labs Individual data in December 2022. We downloaded LinkUp in January 2022. We downloaded Factset data in January 2023. We downloaded CRSP Mutual Fund Holdings Database and Morningstar’s annual Sustainable Funds U.S. Landscape Report in November 2023. We downloaded GlassDoor data in June 2024. We have a Revelio Labs Individual data refresh in December 2024. We downloaded Capital IQ ESG Score data in May 2025. We downloaded Capital IQ Transcripts in July 2024.

Item 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., nondisclosure 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.

All data can be accessed by any subscriber to CRSP Mutual Fund Holdings Database, Morningstar's annual Sustainable Funds U.S. Landscape Report, FactSet, Revelio Labs, Capital IQ ESG Score, Capital IQ Transcripts, Compustat, RavenPack, LinkUp, and Glassdoor. The contact for Revelio Labs is Lisa Simon (lisa@reveliolabs.com), and for LinkUp, the contact is Nate Menard (nate.menard@linkup.com). There are no restrictions imposed by any of the data providers on the publication of results.

Item 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.

We describe data collection and sample construction steps in Section 3 and Appendix A of the manuscript.

Item 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.

All manipulations of the raw data are performed entirely through computer code, using a combination of Python (for data cleaning and construction) and Stata (for statistical analysis and output generation). No manual alterations of the raw data were made outside the code provided.

Item 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.

We share the complete code used to convert the raw data into the final datasets, conduct the analyses, and generate all tables and figures reported in the manuscript. All code is saved in the folder titled "CLM_code," which is inside the zipped folder named "CLM code and log identifiers." Log files for Python code executed in Jupyter Notebook format are included in the same files. Log files for Stata code and Python code executed outside of Jupyter are generated and saved in the "CLM_log" folder.

Those files also serves as a step-by-step guide that enables other researchers to understand and run our code. Finally, we provide the list of identifiers in an excel file called "CLM_identifiers" also in the "CLM code and log identifiers" folder.

Item 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 all log files that document the execution of the entire code in the folder "CLM_log" inside the CLM code and log identifiers" folder.

Item 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.

Dan Li will maintain the data for at least six years.