

Data Description Sheet
for
**Discontinuous Distribution of Test Statistics Around Significance Thresholds in
Empirical Accounting Studies**

Xin Chang, Huasheng Gao, and Wei Li

In this document, we provide additional information regarding the sources of our data and the construction of our samples, as required by the *Journal of Accounting Research* Data and Code Sharing Policy.

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

Wei Li led the empirical works, with ongoing inputs from Xin Chang and Huasheng Gao throughout the process.

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.*

The data were retrieved between 2017 and 2022. Below, we describe how we assembled the final sample by combining the paper and author characteristics with the test statistics reported in those papers.

Paper Dataset.

For the experimental sample, we rely on a Python web scraping script that identifies experimental articles on the journal websites of AOS, CAR, JAE, JAR, RAST, and TAR. We manually check each returned article to ensure accuracy and discard articles that do not report any p -values.

For the archival sample, we randomly select an archival article to match every experimental paper. This archival article needs to be published in the same journal–year as a given experimental article.

Author and Article Characteristics Dataset.

After obtaining the list of articles, we identify the authors of each article and the following characteristics: gender, year of promotion to associate professor, year of obtaining a PhD, affiliated school, PhD school, and the rankings of these schools. These data are hand-collected from authors' school websites, *curriculum vitae*, or social network platforms (e.g., LinkedIn). School rankings are obtained from the UTD Top 100 Business School Research Rankings by counting the number of publications in the

top three accounting journals (TAR, JAR, and JAE). We collect Google Scholar and Web of Science Citation counts from web searches for each article and identify the total number of experimental participants.

Test-statistics Dataset.

To collect the sample of p -values reported in the identified experimental articles, we first convert all articles from PDF format to text format using a Python script and then conduct an automatic keyword search to identify p -values within the text files.

For archival studies, we read each paper to manually identify its main hypothesis. Afterward, we collect the coefficient estimates for the main hypothesis, the test statistics (standard errors or t -statistics, depending on the article's reporting), and the number of observations in regressions.

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).

Not applicable to our paper, as all data used in the paper are publicly available.

4. A complete description of the steps necessary to collect and 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 detailed steps for collecting and processing the data used in the paper are provided in Section 2 of the published paper.

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.

After we retrieved the raw data, all data manipulations were performed using the programs listed in the submission package.

6. *To be provided upon acceptance of the paper and prior to publication: 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 authors have provided a comprehensive replication package that can fully regenerate the paper's results. Interested readers can run the "00_Master_file.do" file to clean the raw data, construct the sample, and generate the tables and figures in our paper. This master do-file invokes the other do-files, numbered according to their execution order in the workflow (e.g., 10 → 11 → 20...).

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.*

The log file is named "log_file.smcl" under the replication package.

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.*

We will maintain all the data and programs for at least six years.