The Silent Majority:
Private U.S. Firms and Financial Reporting Choices

Petro Lisowsky
Boston University Questrom School of Business and
Norwegian Center for Taxation
lisowsky@bu.edu

Michael Minnis
University of Chicago Booth School of Business
michael.minnis@chicagobooth.edu

Compliance with Data Policy for the Journal of Accounting Research

April 2020

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

   Lisowsky was involved in the collection, handling, and analysis of the IRS tax return data. Some supplemental analysis was conducted by Minnis when Sageworks, RMA, or Statistics of U.S. Businesses (SUSB) data were used to compare with the IRS data (e.g., Tables A1A and A1B of the supplemental appendix). Minnis handled the Sageworks, RMA, and SUSB data. Publicly available Compustat data and aggregate IRS Statistics of Income (SOI) data were also used briefly in the study.

2. *A detailed description of how the raw data were obtained or generated, including data sources, the 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.*

   - Our main analyses are based on the IRS tax return data. Petro Lisowsky obtained firm-level tax return data between February and June, 2013, from the Large Business & International Division of the IRS. The data use parent-level consolidated Forms 1120 (C corporations), 1120S (S corporations), and 1065 (partnerships and LLCs). Data from the main portion of the forms were combined with each firm's respective consolidated Schedules M-3 by an IRS data analyst for Lisowsky. The Schedules M-3 are only available for firms with assets of $10 million or more that e-file their tax returns; the relevant GAAP and audit data are located on the Schedules M-3. Lisowsky performed all tax return data analyses.
     - Additional analyses used data from RMA, Sageworks, SUSB, and IRS SOI. These data are only used for benchmarking population statistics, and these data have been

discussed elsewhere. See Minnis (2011) and Lisowsky, Minnis, and Sutherland (2017) for additional descriptions of the datasets (references included in the paper).

- We also use data from Compustat to generate industry-level variables for Table 6. For the final version of the paper, we accessed Compustat via WRDS on January 14, 2015 and our Stata-do files provide all of the code needed to generate our industry-level variables from the raw Compustat data.

Both authors vouch for the stated sources of 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 agreement, any restrictions imposed by the organization on the authors with respect to publishing certain results).*

   We have provided contact information to the editors for the IRS datasets.

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 papers, we require information about subject eligibility and/or selection, as well as any exclusion criteria.*

   We describe our data collection processes in various places throughout the paper for the various datasets used. Our main sample selection procedure is reported in Table 1 of the paper. We also provide a detailed description of our analysis in a separate coding file.

5. *Prior to final acceptance of the paper, the computer program used to convert the raw data into the dataset used in the analysis plus a brief description that enables other researchers to use this program. Instead of the program, researchers can provide a detailed step-by-step description that enables other researchers to arrive at the same dataset used in the analysis. The purpose of this requirement is to facilitate replication and to help other researchers understand in detail how the sample was formed, including the treatment of outliers, Winsorization, truncation, etc. This programming is in most circumstances not proprietary. However, we recognize that some parts of the data generation process may indeed be proprietary or otherwise cannot be made publicly available. In such cases, the authors should inform the editors upon submission, so that the editors can consider an exemption from this requirement.*

   We use Stata to convert the raw data and perform all analyses. Some pre-processing of data (e.g., converting raw data to Stata datasets) occurred prior to using Stata, but the authors did all data manipulation as recorded in the Stata do-files. The only exceptions are steps which are redacted to protect the confidentiality of the underlying data. These steps are minor, but are discussed in the separate coding files provided with this manuscript. Employer Identification Numbers (EIN) are the firm level identifiers we use in the dataset.

6. *Data and programs should be maintained by at least one author (usually the corresponding author) for at least six years, consistent with National Science Foundation guidelines.*

   The IRS data access is limited to two years. The contract between the IRS and Lisowsky stipulates the following clause: "The contractor certifies that the IRS data processed during the performance of this contract shall be completely purged from all data storage components of its computer facility and no output will be retained by the contractor at the time the IRS work is completed. If immediate purging of all storage components is not possible, the contractor certifies that any IRS data remaining on any storage component will be safeguarded to prevent unauthorized disclosures." Lisowsky is continuing to work with the data set on this project, as well as for Barrios, Lisowsky, and Minnis (2019), so data access continues. However, once the other project is completed, the data must be purged. Lisowsky will petition to hold on to the data as long as legally possible.