Data Description Sheet

"Did the Dodd-Frank Whistleblower Provision Deter Accounting Fraud?"

Philip G. Berger

Booth School of Business, University of Chicago

Heemin Lee
Baruch College, City University of New York

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

Heemin Lee has handled all data and conducted the analyses. Data handling and analyses performed are described in more detail below.

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 used for all analyses in the paper were downloaded from Wharton Research Data Services (WRDS). The dataset includes Thomson Reuters 13F institutional holdings filings (downloaded on 7/2/2016), Compustat (downloaded on 6/20/2016), audit fees (downloaded on 7/21/2016) and internal control weaknesses (downloaded on 7/26/2016) from Audit Analytics, and Compustat business segment data (downloaded on 8/1/2016).

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).

All data used in this paper came from publicly available sources.

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.

We describe our sample selection process in section 4 and Table 1 of the paper. First, we downloaded the raw data from Thomson Reuters 13F institutional holdings filings, Compustat, and Audit Analytics dataset through WRDS for a sample period of 2007 – 2014 (specifically, we use data from 2008 – 2014 for the Dodd-Frank whistleblower law tests and from 2007 – 2010 for state False Claims Act (FCA) tests). Second, we identified 22 state pension funds and collected their state FCA information from Bucy et al. (2010) and Rapp (2012b) (the lists of funds and state FCAs information are provided in Table 2 and Appendix B, respectively). Third, we matched these state-level pension fund and FCA data with 13F portfolio holdings, Compustat, and Audit Analytics data of U.S. public firms. For firms with shares owned by multiple state pension funds, we aggregate their ownership information at the firm-year level. The firm-years of public Compustat firms not matched with 13F filings are also kept in the sample. Fourth, we excluded the healthcare and financial industries. Fifth, after eliminating missing values, we required firms to have at least one observation from both pre- and post-Dodd-Frank periods for the Dodd-Frank analysis. For the state FCA analysis, we required firms to have more than one observation during 2007 – 2010. Lastly, we winsorized observations at the bottom and top 1% for continuous variables.

5. The computer programs or code used to convert the raw data into the final dataset used in the analysis plus a brief description that enables other researchers to use this program. 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, 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 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 final dataset used in the analysis. In such cases, the authors should inform the editors upon initial submission, so that the editors can consider an exemption from the code sharing requirement. Whenever feasible, authors should also 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 used Stata to process the raw data, define variables, construct the final datasets, and perform the main analyses. These steps are described in the program file "*BL_Final_Code.do*" that we provide to fulfill JAR's Data and Code Sharing Policy. We also include the file "*BL_Identifiers*" that lists CUSIP identifiers for the firms in our final sample for the Dodd-Frank analysis and the state FCA analysis.

6. 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 authors agree to maintain the data and programs for at least six-years, consistent with National Science Foundation guidelines.