

## **Data Description Sheet: “Fraud Allegations and Government Contracting” by Jonas Heese and Gerardo Pérez Cavazos**

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

Data was handled and analyzed by Jonas Heese and Gerardo Pérez Cavazos.

**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 is able to vouch for the stated source of the raw data.**

#### **FCA Lawsuits**

Data on whistleblower events was obtained from Professor David F. Engstrom (Professor of Law at Stanford University) via Email on July 23, 2015. Engstrom obtained the data through a series of requests to the DOJ under the Freedom of Information Act (see Engstrom [2013] for details). His data included the following case information: the date the lawsuit was received by the court, the court docket number of the lawsuit, the caption of the lawsuit, the judicial district where the lawsuit was filed, the primary agency alleged of being defrauded, the DOJ’s election decision and election date, and the settlement date as well as amounts. We then download the court filing for each lawsuit from the Public Access to Court Electronic Records (PACER) system. We manually review the court docket filing associated with each case and match the name of every publicly listed defendant to a GVKEY identifier. Specifically, we first apply a fuzzy-match approach to match the names of the defendant firms to Compustat company names. We then manually review this matching procedure and only keep correctly identified pairs. The paper describes how the final sample is formed (see Table 1, panel A).

As described in Appendix B, we create the following variables from this database: *Settlement*, *DOJ\_Intervention*, *Long\_Investigation*.

#### **Government Contracts**

We obtain information on government contracts from the USAspending.gov website maintained by the Department of Treasury’s Bureau of Fiscal Services. We downloaded the data in February 2017. The database does not contain common identifiers such as GVKEY or PERMNO, so we use a machine-learning program to fuzzy-match over 596,000 parent company names to the names of our whistleblower companies, then hand-check identified pairs to ensure correct matching.

#### **Political Connections**

We obtain data on firms’ Political Action Committee (PAC) contributions from the Federal Election Commission’s (FEC) website. As this database does not contain common identifiers such as GVKEY or PERMNO, we again fuzzy match the names of the PACs to the names of our whistleblower companies. Finally, we hand-check identified pairs to ensure correct matching.

We obtain data on firms’ lobbying expenditures from the Center for Responsive Politics (CRP). Again, we fuzzy match the company names of the CRP database to the names of our whistleblower companies. Finally, we hand-check identified pairs to ensure correct matching.

We obtain data on political connections of firms’ board members from firms’ Def14a filings. If (at least) one of a firm’s board members has been president, (vice-)presidential candidate, senator,

congressman, governor, mayor, chairman of the Party Caucus, member of the president's cabinet, White House staff, or campaign staff, we define the board as politically connected.

All other variables are created using commonly used databases such as Compustat, and are described in Appendix B.

**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, any restrictions imposed by the organization on the authors, such as restrictions to publish certain results).**

The data was not obtained from an organization on a proprietary basis.

**Item 4. A complete description of the steps necessary to collect and process the data used in the final analyses reported in the paper.**

The steps necessary to process the data used in the final analyses are described in the Data Section of the paper. Appendix B provides a description of the variables used in the empirical analyses. Further detail of the steps taken to collect the data and the corresponding code can be found in the files submitted to the *Journal of Accounting Research*. The data were processed using a mix of SAS and Stata and the primary analyses were performed in Stata.

**Item 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. The 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.**

Code files are provided and include comments throughout. A list of identifiers (GVKEY) for the final sample can be found in the Identifiers.txt file included with the submission.

**Item 6. An assurance that the data and programs will be maintained by at least one author for at least six years, consistent with National Science Foundation guidelines.**

The authors agree to maintain the data and programs used in this paper for the six-year time period suggested by the National Science Foundation.