Data Description Sheet: "What are you Saying? Using *topic* to Detect Financial Misreporting" by Nerissa C. Brown, Richard M. Crowley, and W. Brooke Elliott (2020)

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

Richard Crowley: Collected all SEC EDGAR data and conducted all analysis except the computation of the quantitative financial statement and stock market measures.

Nerissa Brown: Collected all Compustat, CRSP, AAER, and Audit Analytics data, and generated the quantitative financial statement and stock market measures used in our models. Analyzed participant responses from the human-subject word intrusion task conducted on Amazon Mechanical Turk (MTurk).

W. Brooke Elliott: Served as principal investigator for the word intrusion task conducted on Amazon Mechanical Turk. Nerissa Brown, Richard Crowley, and Brooke Elliott developed the experimental instrument used in the task.

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.

SEC EDGAR data was obtained over the span of two weeks starting in October of 2013 from the SEC EDGAR FTP site (publicly available). In order to determine the relevant files, all index files were first downloaded from the FTP site and parsed using a VBA for Excel script that located all 10-K and 10-K/A filings from 1994 through 2012. Then, the GNU Wget computer program was used to systematically download all 10-K and 10-K/A filings from the FTP site.

All the restatements identified from 10-K/A filings were gathered using Python scripts developed by Richard Crowley. All textual style measures in the study are computed using Python scripts developed by Richard Crowley. All the LDA topic measures are computed using a combination of Python scripts written by Richard Crowley and a version of onlineldavb (python script, available on github) with edits made by Richard Crowley.

Financial statement and stock market variables were downloaded from Compustat and CRSP via the Wharton Research Data Services (WRDS) platform in October of 2013. The Audit Analytics restatement variables were downloaded via the WRDS platform in July 2018. The SEC AAER data were obtained in July 2018 from the Center for Financial Reporting and Management (CFRM) at the University of California – Berkeley. The quantitative financial statement and stock market measures used in the *F-score* model are computed using Stata scripts developed by Nerissa Brown. All other statistics and analysis are computed using Python scripts developed by Richard Crowley, using Pandas and statsmodels as the statistics backend. Further analysis is done in IPython notebooks and Stata *.do files developed by Richard Crowley.

The human-subjects word intrusion task was administered on MTurk in December 2017. Participants responses were gathered using Qualtrics and analyzed using Stata scripts developed by Nerissa Brown.

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).

The SEC AAER data was purchased in July 2018 from the Center for Financial Reporting and Management (CFRM) at the University of California - Berkeley. The dataset extends the AAER sample used in Dechow et al. [2011] and contains all AAERs issued between May 17, 1982 and September 30, 2016. The data is collected from the SEC's publicly-available repository of enforcement actions, available at http://www.sec.gov/divisions/enforce/friactions.shtml. The SEC AAER dataset is now being housed at the University of Southern California's Leventhal School of Accounting. Additional information about this dataset is available at https://sites.google.com/usc.edu/aaerdataset/home.

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.

A complete description of the parsing and processing of the 10-K filings is available in Appendix A.1 of the online appendix. Details on data processing are available in Section 3 ("Data and Empirical Methodologies") of the paper, and we provide additional details and code as outlined under item #5 below.

Participants were recruited on MTurk for completion of the human-subjects word intrusion task. The subject pool was restricted to individuals who are U.S. citizens, proficient in English, and are a least 18 years of age. We limit participation to U.S. citizens due to policies at the researchers' universities that impose restrictions on payments to non-U.S. citizens. We determine English proficiency by asking participants to indicate whether English is their native language. Our original sample consists of 200 participants. We eliminate 12 responses with duplicate IP addresses and 2 responses that indicate English as a non-native language. We also exclude 6 participants who completed the task in less than 60 seconds to minimize concerns about participant effort in online labor markets (Farrell, Grenier, and Leiby [2017]). These exclusions lead to a final number of 180 participants.

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.

All code in the "Code_transfer.zip" file will replicate the LDA topic measure constructed in Brown, Crowley, and Elliott (2020). There is a "README.txt" file which contains instructions on how to get the code up and running, and a "requirements.txt" file which contains the specific python packages, with their versions, which are sufficient to execute the code. Since there are a number of python scripts provided in the zip file, a bash script to execute all needed scripts in order is provided as "Z_log/main.sh".

The comma-delimited (CSV) and Stata (DTA) files contain firm identifiers (GVKEYs and CIKs) and variables for the restatements collected from amended 10-K filings (10-K/As) using the text script detailed in Appendix A of the paper. These restatements are referred to as "10-K/A Irregularities" in the paper. Restatements that are not classified as irregularities can be identified from the "restate_ben" variable. Brief definitions of each variable are as follows:

- **gvkey**: Compustat firm identifier
- **cik**: Central Index Key
- **year**: calendar year
- **datefiled**: Filing date of original 10-K filing in YYYY-MM-DD format
- **fye**: fiscal year end date of original 10-K filing in MM/DD/YYYY format
- **restate_filing**: SEC-assigned accession number for the 10-K/A filing
- **daterestated**: Filing date of the Amended 10-K filing in YYYYMMDD format
- **restatement**: A binary indicator to identify a restatement of a firm-year's 10-K. Equals 1 if restate_filing is not blank, 0 otherwise.
- **restate_dir**: A binary indicator to identify direct irregularity restatements (direct mention of fraud, an irregularity, materially false and misleading information, or a violation of securities law in the 10-K/A filing)
- **restate_govt**: A binary indicator to identify irregularity restatements identified by government or regulatory entities, namely the SEC, Department of Justice (DOJ), and/or the Office of the Attorney General.
- **restate_oth**: A binary indicator to identify irregularity restatements identified by third party investigations, such as those identified by forensic investigations or audit committees, or those involving legal counsel.

- **restate_int**: Equals 1 if restate_dir, restate_govt, or restate_oth are equal to 1. Equals zero if all three of these indicator variables are equal to zero.
- **restate_ben**: Equals 1 for all other restatements that are not classified as an irregularity (i.e., equals 1 if restate_int is equal to zero)

Please refer to Appendix A of the paper for a full description of the text script used to identify irregularities from amended 10-K filings.

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.

Nerissa Brown and Richard Crowley will maintain all the data and programs used until at least 2026.