

Journal of Accounting Research Data Policy

Data Description Sheet to **Why do individual investors disregard accounting information?** **The roles of information awareness and acquisition costs**

Elizabeth Blankespoor, Ed deHaan, John Wertz, and Christina Zhu

September 2018

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

All authors had access to the raw index of AP articles. Christina Zhu downloaded additional data sources and matched the index of articles to the company information. Elizabeth Blankespoor analyzed the AP article text to identify various components included. Both Christina Zhu and John Wertz handled further data sources and conducted analyses. All authors examined the analyses and have access via a shared Dropbox folder to the computer programs used to perform the analyses.

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 sources and download dates are as follows:

Data	Source	Download Date
Associated Press (AP) automated earnings articles index	AP	11/20/2015
AP automated earnings articles text	Yahoo and AP webpages	February 2016
Compustat Fundamentals Annual	WRDS	3/25/2016
Compustat Fundamentals Quarterly	WRDS	2/28/2017
Compustat Header File	WRDS	9/12/2017
CRSP Daily Stock	WRDS	10/4/2016; 9/27/2017
CRSP Monthly Stock	WRDS	10/4/2016; 9/27/2017
CRSP Stocknames File	WRDS	9/26/2017
IBES Actuals and Estimates	WRDS	6/14/2017 – 6/16/2017
IBES Guidance	WRDS	9/27/2017

NYSE Trade and Quote (TAQ)	WRDS	March 2016; February 2018
Ravenpack	Ravenpack	October 2015; April 2016
WhaleWisdom Institutional Ownership 13F Filings	WhaleWisdom	5/10/2017
Wall Street Horizon Earnings Announcement Dates	Wall Street Horizon	8/29/2016
Zacks Actuals and Estimates	WRDS	6/15/2017
Zacks Preannouncement	WRDS	1/16/2017

NYSE TAQ: Christina Zhu and John Wertz downloaded and compiled TAQ millisecond data during March 2016 and February 2018. The process took a few weeks due to the size of the data. Elizabeth Blankespoor, Ed deHaan, John Wertz, and Christina Zhu can vouch for this source.

Wall Street Horizon: Ed deHaan obtained data from Wall Street Horizon. Ed deHaan and Barry Star, CEO of Wall Street Horizon, can vouch for this source.

WhaleWisdom: Christina Zhu downloaded and compiled data from WhaleWisdom's 13F Filings in May 2017. Christina Zhu can vouch for this source.

AP automated articles: We obtained the list of AP articles from AP, and we scraped the article text from Yahoo and AP webpages. Elizabeth Blankespoor, Ed deHaan, and Christina Zhu can vouch for the sources.

All authors can vouch for the remaining data downloaded from WRDS.

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).

Proprietary data on AP's news articles were provided by AP under a Non-Disclosure and Confidentiality Agreement signed by Ed deHaan. We have privately provided the editors with a copy of the Agreement and the contact information for a company representative from AP. All other data are publicly available from the sources listed above.

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 paper verbally describes the data collection and processing steps. Further detail of the steps taken to collect the data and the corresponding code can be found in the 0_Overview.txt file submitted to the *Journal of Accounting Research*. We obtained the data outlined in our response to Question 2 above using SAS and

a mix of remote submits and direct requests on the WRDS servers. The data were processed using a mix of SAS and Stata and the primary analyses were performed in Stata.

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.

Computer programs are outlined in the 0_Overview.txt file. The code files are sequentially ordered and a brief description of each file is included in the Overview. Code files include comments throughout. A list of identifiers (GVKEY and DATADATE) for the final sample of 29,776 observations can be found in the Identifiers.txt file included with the submission. As in our initial submission, the code necessary to match and clean the raw AP data are withheld as proprietary. See the draft for details on the process as well.

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 used in this paper for the six-year time period suggested by the National Science Foundation.