Nicholas Guest
Assistant Professor of Accounting

314 Sage Hall
Ithaca, NY 14853-6201

L. 607.255.4220
C. 607.279.5990
E. nguest@cornell.edu

Samuel Curtis Johnson
Graduate School of Management
Cornell University

Johnson
Cornell
SC Johnson College of Business

Data Description for "The Information Role of the Media in Earnings News"

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

The sole author of the paper, Nick Guest, conducted the data collection and 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.

Much of the data for analyses in the paper were downloaded from Wharton Research Data Services (WRDS) using remote submitting (i.e., using the SAS command "rsubmit"). The central dataset was downloaded on October 17, 2017. The datasets used are the following: CRSP Daily and Monthly Stock, Compustat Annual and Quarterly Fundamentals, IBES Consensus, SEC Analytics, RavenPack Dow Jones, Thomson Reuters 13f, and Trade and Quote (TAQ).

I downloaded WSJ article data and text from Factiva.com in October and November 2016. I downloaded earnings press release data and text from SEC EDGAR in December 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).

The data were not obtained on a proprietary basis.

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.

See Sections 3, 4, and Appendix B in the study, as well as the Internet Appendix.

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.

The computer programs used to generate the final dataset, a brief description of the programs, and the primary identifiers for the final sample, have been included with the final submission. A brief description of each file is also included here for convenience, as follows:

1CollectMediaCoverage.py: Python script used to collect the text of WSJ articles from Factiva

2AnalyzeMediaCoverage.py: Python script used to clean article text and calculate linguistic variables used in the study (e.g., percentage of negative words and sentences with quotes)

3CollectPressReleases.py: Python script used to collect the text of firms' earnings press releases from SEC EDGAR

4AnalyzePressReleases.py: Python script used to clean press release text and calculate complexity variables used in the study (e.g., readability and specificity)

5CompareMediaCoverageAndPressReleases.py: Python script used to compare earnings press releases and the associated WSJ articles via the Jaccard (1901) similarity measure

Guest_MediaEarnings_FinalDatasetAndAnalyses.sas: SAS program used to combine the raw outputs from the above mentioned Python programs with other financial and accounting data (e.g., Compustat and CRSP). That is, this code converts raw data into the final dataset used in the study. This file also includes code that produces the key tables and figures in the paper.

identifiers.xlsx: Excel file that provides the gvkey, cusip, and permno for the 494 unique S&P 500 firms that are used as either a treatment or control in the main analyses of the paper.

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.

Data and programs will be maintained by the author for at least six years.