PROJECT INFORMATION LITERACY

News Study Dataset

October 2018

User Guide

Principal Investigator
Alison J. Head, Ph.D.

Co-Investigators
John Wihbey
P. Takis Metaxas, Ph.D.
Margy MacMillan
Dan Cohen, Ph.D.

PIL Survey Team
Erica DeFrain, Ph.D.
Kirsten Hostetler
Elizabeth Berman (Fellow)
Bridget Peery (Fellow)

User Guide: Project Information Literacy's News Study Dataset

Table of Contents

The following APA citation can be used for referencing the PIL news study dataset:

> Head, Alison J., DeFrain, Erica L., Hostetler, Kirsten. (October 16, 2018). *News Study Dataset*. Project Information Literacy. [Data file]. Retrieved from https://doi.org/10.17760/D20293517

1. About the News Study Dataset

The Project Information Literacy (PIL) news study dataset was produced as part of a yearlong scholarly research study funded by the Knight Foundation and the Association of College and Research Libraries (ACRL). This study examined how today's young adults find and engage with news.[1] The goal of the survey was to collect data about the news-seeking behaviors of a sample of students from 17 U.S. institutions, comprised of 11 colleges, universities, and community colleges, and six high schools.

The dataset contains responses from 6,049 total respondents to a 20-item questionnaire administered between February 12, 2018 and April 21, 2018, and the dataset includes responses to a total of 17 of the 20 total questions from the survey. Of these questions, 12 items focused on the news habits during the previous week, including news sources consulted, news topics followed, news items shared on social media, and evaluation practices, and opinions about the role of news in respondents' lives. Another five questions collected demographic data. The last three questions asked whether participants were interested in a follow-up interview, sharing their Twitter screen names with researchers, if they had one, and entering a drawing for a $150 Amazon e-gift card with one given to a winner from each school in the sample.

The voluntary sample of respondents consisted of young adults at least 18 years of age, currently registered as full-time undergraduate students in one of the 11 postsecondary institutions (N = 5,844). There was also an exploratory sample of seniors in six high schools (N = 205). Institutions were selected based on their regional diversity, demographic variation, and whether they were located in red or blue states.[2]

The dataset includes quantitative data about how young adults interact with and consume news in their personal and academic lives. Demographic information for the respondents — including age, gender, class standing, major, and political affiliation — is also included in the dataset.

A 53-page research report, "How Students Engage with News," was published on the PIL site on October 16, 2018. A copy of the report and dataset was also submitted to the Knight Foundation and the Association of College and Research Libraries, the two funders for the research for uploading on their websites as well as Gutman Library at Harvard Graduate School of Education and Snell Library at Northeastern University. This report is an open access document and includes a Methods section. Links to the Survey Instrument and other related materials are on the site.

2. Research Questions

The study examined young adults' news engagement habits and preferences. The survey was used to collect data about the topics followed and sharing behaviors of young adults, and about the news sources on which they relied. For purposes of our study, news was defined as information about events happening all around the world. News comes to us through a variety of social, digital, and physical channels.

Three questions guided our research study:

1. How do students conceptualize what constitutes "news" and how do they keep up, if they can?
2. How do students interact with and experience news when using social media networks?
3. How do students determine the currency, authority, and credibility of the news content they encounter from both traditional and new media sources?

The study was one of the largest research efforts of its kind, asking questions that go beyond simple news preferences to understanding the news habits and opinions about the role of news in students' lives. Findings are intended to help better inform public discourse, potential policy solutions, and courses of action for librarians, journalists, and educators as they grapple with critical issues of misinformation and credibility in the media environment.

[1] The John S. and James L. Knight Foundation and the Association of College and Research Libraries (ACRL), the largest division of the American Library Association, funded this study with support from Northeastern University's Snell Library and the College of Arts and Media Design (CAMD), and Harvard Graduate School of Education (HGSE). Dr. Alison J. Head, the founder and director of PIL, was the study's Principal Investigator. Communication about this report should be sent to Dr. Alison J. Head at alison@projectinfolit.org.

[2] The institutional sample was made up of the following 17 US colleges, universities, and high schools: Belmont University (TN), Brandeis University (MA), California State University Maritime Academy (CA), DePaul University (IL), John Tyler Community College (VA), Oklahoma State University (OK), Saint Mary's College of California (CA), University of Alaska Anchorage (AK), University of Michigan (MI), University of Texas Austin (TX), Wellesley College (MA), Benicia High School (CA), Dayton Regional STEM School (OH), Fergus High School (MT), Mercy High School (MD), Yellow Springs High School (OH), and Wellesley High School (MA).

TABLE 1: OVERVIEW OF DATA COLLECTION PHASES

Research Phase	Dates	Summary of Research Activities
Phase One: Online Survey	Feb. 12, 2018 – April 21, 2018	A 20-item survey was completed by a sample of 6,049 currently registered undergraduate and high school students, who were18-years-old or older. The survey was used for collecting data about the news engagement behaviors of a sample of students who volunteered to participate in the study.
Phase Two: Twitter Analysis	Jan. 10, 2018 – May 31, 2018	A computational analysis was conducted comparing data from 765 Twitter screen names from college and high school survey respondents in this study with an existing large-scale panel of 135,891 Twitter users.
Phase Three: Follow-up Interviews	May 4, 2018 – June 15, 2018	A three-item, open-ended telephone interview script was used in follow-up telephone interviews with a sample of 41 high school and college participants for collecting qualitative data. The sample subset volunteered for a follow-up interview to the survey.

3. Study Design

The study investigated students' news engagement practices, attitudes, and preferences. The dependent variable was the behaviors employed by members of the sample to meet their news gathering needs. These behaviors included, but were not limited to, following news on such topics as national government and politics, traffic and weather, business, crime, race and immigration, schools and education, etc., and evaluating and sharing news.

The independent variables were the pathways to news students used and a constellation of related demographic variables (e.g., college major, age, gender, and political affiliation). Pathways to news were, but are not limited to, social media, print newspapers, online newspaper sites, television, radio, podcasts, discussions with peers, discussions with professors, and discussions with librarians. Attitudes included considerations of the role and credibility of news and the impact of 'fake news'. Preferences included news that was useful in students' lives, drew their interest, included objective reporting, etc.

4. Sample Design

4.1 Institutional Sample

The institutional sample consisted of 11 U.S. postsecondary institutions (i.e., one community college, five private colleges and universities, and five public colleges and universities), and six U.S, high schools. Institutions were selected from PIL's volunteer sample of over 265 institutions to establish a diverse geographic and demographic representation of young adults in America.

4.2 Survey Sample

The voluntary sample for the survey was drawn from the representative population and surveyed one time. To qualify for taking the survey, a respondent must (1) have been 18 years of age or older, and (2) currently registered as a full-time student at one of the 17 institutions.

5. Data Collection Activities

Data collection for the study took place in three phases (Table 1). An overview of the phases and associated research activities appears in Table 1. In Phase One, an online survey was administered to students at 17 different high schools and colleges and universities. In Phase Two, a computational analysis was conducted of Twitter screen names from survey respondents who agreed to share account information. In Phase Three, semi-structured telephone interviews were used to collect qualitative data from a sample of 41 students who completed the survey.

Only the quantitative data from Phase One — the online survey —has been made available in the public dataset. Data from Phases Two and Three are not available in the public dataset.

Prior to data collection, the PIL research protocol underwent Human Subjects Division Review at Wellesley College, where Panagiotis "Takis" Metaxas, a co-researcher on this study, is a professor of computer science. The research protocol also obtained approval from the 17 postsecondary and high schools in our institutional sample.[3]

[3] The Wellesley IRB was approved on December 21, 2018. The Federal-wide Assurance (FWA) #: 00000598 and the IRB Registration #: 00001401.

6. Data Collection: The Survey

6.1 Sample Recruitment

Survey data were collected from respondents between February 12, 2018 and April 21, 2018. PIL collaborated with staff at each institution in the study, which either deployed the survey or provided contact information for the participant sample. Email invitations for study participation were sent to all currently registered undergraduate and high school students aged 18 years or older that made their email addresses publicly available. At institutions with large enrollments (and thus having a larger number of potential participants), a random subset sample of eligible respondents was used. At institutions with smaller enrollments, a voluntary sample of eligible respondents was used to ensure a certain number of responses.

The email invitation asked students to volunteer for the study by taking an online survey, and a link to the survey at Wellesley College was provided. The survey was described to potential respondents as being about their experiences interacting with and consuming news and gathering information.

6.2 Mode of Data Collection

Survey data were collected using the Qualtrics survey application licensed through Wellesley College, where one of the co-researchers is a professor. Survey results datasets were downloadable in both SPSS and comma-delineated (CSV) formats for analysis.

6.3 Survey Questions

The public dataset includes responses to a total of 17 of the 20 total questions from the high school and college surveys. Of these, 11 items focused on the news engagement habits during the past week, including news sources consulted, news topics followed, news items shared on social media, and evaluation practices, and opinions about the role of news in respondents' lives. Another six questions collected demographic data.

Five questions that had the potential to disclose personal identifiers were altered or omitted from the public dataset: Question 12 (open-ended question about conducting research for fulfilling assignments vs. personal use); Question 13 (the institution where a student was enrolled); Question 19 (sharing Twitter screen name); Question 20 (signing up for a follow-up interview); and an unnumbered item (entering a drawing at each institution for a $150 Amazon e-gift card).

College students were asked about their major field of study in Question 15, while high school students were asked if they planned to attend college the next year in Question 15. On average, completing the survey took 12 - 15 minutes. To enhance the reliability of this study's survey results, the survey instrument was pilot tested with 11 students matching our selection criteria but who were not eligible for the study sample. An overview of the survey questions and the variable names used in the dataset appears in Table 2.

6.4 Response Rate

To qualify for the survey, a respondent must have been 18 years of age or older and currently enrolled as a full-time student in a college or high school in our institutional sample. In all, 60,541 college students and 250 high school students were invited to take the online survey. Of 6,049 respondents that clicked on the survey link, the large majority were college students (N = 5,844), with a subset of high school students 18 years or older (N = 205). Response rates varied per institution, with an overall response rate of 9.65%.

7. Data Preparation: Public Dataset

Data in the public dataset are presented as an aggregate of all 17 institutions in the sample, i.e., college and universities and high schools. All personal identifiers – names and email addresses for scheduling a follow-up telephone interview and for entry into an Amazon gift card contest, and Twitter handles – were removed from the dataset before data analysis in order to protect the privacy of study participants. Missing data were imputed as Refusal/ No Answer.

The dataset was saved in one downloadable format: CSV. The survey dataset includes responses from the entire sample, 6,049 students of 17 US postsecondary and high schools. The dataset file also includes the user guide (PDF), the survey instrument (PDF), and a codebook with topline data response per question (PDF).

TABLE 2: NEWS ENGAGEMENT SURVEY
(Questions available in the public dataset)

Survey Question	Variable name(s) in Survey Dataset	Survey Question(s) Description
Q1: How often, if at all, has your news come from one of these sources during the past week?	Q1_SocialMedia; Q1_NewsFeeds; Q1_Print; Q1_Online; Q1_Television; Q1_Radio; Q1_Podcasts; Q1_Peers; Q1_Teachers; Q1_Librarians	A matrix question was asked about how often during the past week the respondent had received news from 10 different pathways to news. A frequency scale was used for responses with 5 options ranging from "Didn't use this source at all this week" to "Several times a day."
Q2: In the past week, how often have you read, listened to, or viewed news items about the following topics, given what's going on in the world around you now?	Q2_TrafficWeather; Q2_Environment; Q2_Local; Q2_NationalPoli; Q2_Business; Q2_Crime; Q2_International; Q2_Health; Q2_Memes; Q2_Education; Q2_Science; Q2_Race; Q2_Sports; Q2_Entertainment; Q2_Lifestyle; Q2_Art	A matrix question was asked about how often during the past week the respondent had consumed news according to 16 different topics. A frequency scale was used for responses with 5 options ranging from "Didn't follow this news at all this week" to "Several times a day."
Q3: If you had only one of these five sources available to you this week, which one would you choose for getting news about the U.S. national government and politics?	Q3_OneSource	A multiple-choice question was asked about which source a respondent would use to get U.S. government and political news if they could only choose 1 during the week. A "Click ONLY one" response format was used.
Q4: How often, if at all, did your news come from one of these social media sites during the past week?	Q4_Facebook; Q4_Instagram; Q4_LinkedIn; Q4_Pinterest; Q4_Reddit; Q4_Snapchat; Q4_Tumblr; Q4_Twitter; Q4_YouTube	A matrix question was asked about how often during the past week the respondent had received news from 9 social media sites. A frequency scale was used for responses with 6 options ranging from "I don't use this social media source at all" to "Several times a day."

CONTINUED >

Survey Question	Variable name(s) in Survey Dataset	Survey Question(s) Description
Q5: How often, if at all, have you shared or retweeted a news item on the social media sites that you use about one of these topics during the past week?	Q5_ShareTraffic; Q5_ShareEnviro; Q5_ShareLocal; Q5_ShareNatlPolitics; Q5_ShareBusiness; Q5_ShareCrime; Q5_ShareIntl; Q5_ShareHealth; Q5_ShareMemes; Q5_ShareEduc; Q5_ShareScience; Q5_ShareRace; Q5_ShareSports; Q5_ShareEntert; Q5_ShareLifestyle; Q5_ShareArt	A matrix question was asked about how often during the past week the respondent had shared a news item via social media according to 16 different topics. A frequency scale was used for responses with 6 options ranging from “I don’t share or retweet news items about this” to “Several times a day.”
Q6: Why do you share news items, if at all, on the social media sites that you use? Please indicate how strongly you agree or disagree with each of the following statements.	Q6_DefineOnlinePresence; Q6_SomethingOthersShouldKnow; Q6_ProvokeResponse; Q6_EntertainFriends; Q6_EntertainMyself; Q6_GiveVoiceAbtCause; Q6_ChangeOthersViews; Q6_TakeaBreak	A matrix question was asked about why respondents share news items via social media according to 8 possible reasons. A 6-point Likert scale was used to capture level of agreement, ranging from “Strongly disagree” to “Strongly agree,” including the statement “I don’t share or retweet news items at all.”
Q7: When you're deciding to share "breaking news"--a special news event that is currently developing--on social media how do you evaluate the quality of the information that you share, if you do at all?	Q7_CheckCurrency; Q7_CheckHashtag; Q7_CheckWhoPosted; Q7_CheckOrigins; Q7_FactCheck; Q7_ReadComments; Q7_TimesLiked; Q7_TimesShared; Q7_ReadEntireStory; Q7_TakeScreenShot; Q7_GoWithGut	A matrix question was asked about how often during the past week the respondent had shared a news item via social media according to 11 different topics. A frequency scale was used for responses with 6 options ranging from “I don’t share breaking news at all” to “Almost always.”
Q8: News can be defined in different ways depending on your point of view. In this question, we want to learn what news means to you and the role that news plays in your life. From your perspective, how much do you agree or disagree with the following statements about what constitutes news, whether it comes from social media feeds, news sites, or print sources?	Q8_Useful; Q8_Factual; Q8_Helps2UnderstandWorld; Q8_ObjectiveReporting; Q8_Necessary4Democracy; Q8_CivicResponsib; Q8_Hard2Tell; Q8_Difficult2TellFakeNews; Q8_Overwhelming; Q8_DoesntThink	In a matrix question, respondents were asked to express their level of agreement with statements about news. A 6-point Likert scale was used to capture level of agreement, ranging from “Strongly disagree” to “Strongly agree,” including the statement “I don’t know.”

CONTINUED >

Survey Question	Variable name(s) in Survey Dataset	Survey Question(s) Description
Q9: News stories sometimes contain factual errors. From your perspective, how much do you agree or disagree with the following statements about news stories, and the journalists that produce them? - From my perspective...	Q9_NoTrust; Q9_FakeNewsImpact; Q9_NoOriginNoTrust; Q9_TrustProfJourns; Q9_JournsInsertBias; Q9_JournsMakeMistakes	In a matrix question, respondents were asked to express their level of agreement with 6 statements about news and journalism. A 6-point Likert scale was used to capture level of agreement, ranging from "Strongly disagree" to "Strongly agree," including the statement "I don't know."
Q10: "Fake news" is a term we hear and see a lot these days. How confident do you feel with recognizing fake news?	Q10_RecognizeFake	A multiple-choice question was asked about how confident a respondent felt in their ability to detect fake news. A 6-point Likert scale was used to capture level of confidence, ranging from "Not confident at all" to "Very confident," including the statement "I don't know."
Q11: Now we'd like to ask you about something different. Some, but not all, students say there are differences between how they get news for fulfilling academic assignments vs. how they get news for personal use in their lives. How do you get news for fulfilling academic assignments vs. news for personal use in your life?	Q11_academic_LibraryDatabases; Q11_personal_LibraryDatabases; Q11_academic_Nontrad; Q11_personal_Nontrad; Q11_academic_teacher; Q11_personal_teacher; Q11_academic_SocialMedia; Q11_personal_SocialMedia; Q11_academic_Apps; Q11_personal_Apps; Q11_academic_Print; Q11_personal_Print; Q11_academic_Television; Q11_personal_Television; Q11_academic_Radio; Q11_personal_Radio; Q11_academic_DontRely; Q11_personal_DontRely	Respondents were asked to identify whether they used 9 different news sources for their academic assignments or personal lives. A category for "I don't rely much on the news" was included as the last option for this question.
Q13: Where are you currently enrolled as a student?	Q13_School	Respondents were asked to identify their institution. Per IRB, data containing individual institutional identifiers were replaced to identify a respondent only as being from either the high school or college sample.
Q14: What's your current status as a student (e.g., first-year, sophomore, junior, or senior, or otherwise)?	Q14_SchoolYear	Respondents were asked to identify their status as a student. A category for "other" was included for this question.
Q15a: Are you planning to attend college next fall? (i.e., high school sample only asked this question)	Q15a_WillAttend	The survey for high school participants asked whether they would be attending college. A category for "other" was included along with a write-in response box.

CONTINUED >

Survey Question	Variable name(s) in Survey Dataset	Survey Question(s) Description
Q15b: What is your major area of study? (i.e., college student sample only asked this question)	Q15b_Major	The survey for college participants asked about their undergraduate major according to 11 different options, including "other," which was included as the last category for this question. A "Click ALL that apply" response format was used.
Q16: What is your age today?	Q16_Age	Respondents were asked to identify their current age according to 4 possible age ranges. A category for "Prefer not to state" was included as the last category for this question.
Q17: What do you identify yourself as?	Q17_Gender	A question was asked about the respondent's gender. A category for "Prefer not to state" was included as the last category for this question.
Q18: How do you describe yourself politically?	Q18_Politics	A question was asked about the respondent's political identity. A category for "Prefer not to state" was included as the last category for this question.

8. Intellectual Property Rights

The intellectual property rights to the survey dataset have been determined as a condition set forth in the ACRL and Knight Foundation funding agreements. Both parties firmly agree that the dataset must be open access and freely available for (1) research, scholarly or academic purposes, or (2) a user's own personal, non-commercial use.

8.1 Licensing

The dataset has a Creative Commons (CC) license of "CC BY-NC-SA 4.0." This license allows others to share, copy, adapt, and build upon the survey data non-commercially, as long as the source – Project Information Literacy – is credited and users license their new creations under the identical terms.

Users of the dataset may not reproduce, sell, rent, lease, loan, distribute or sublicense, or otherwise transfer any of the data from the PIL survey dataset in whole or in part, to any other party, or use the data to create any derivative work or product for resale, lease or license. Nonetheless, a dataset user may incorporate limited portions of the data in scholarly research or academic publications or for the purposes of news reporting, provided that the user of the dataset references the source.

8.2 Citation Format

The following APA citation can be used for referencing the PIL News Study Dataset:

> Head, Alison J., DeFrain, Erica L., Hostetler, Kirsten. (October 16, 2018). *News Study Dataset*. Project Information Literacy. [Data file]. Retrieved from https://doi.org/10.17760/D20293517

The survey sample of college and high school students includes self-selected study participants. The sample was not fully randomized and is not representative of the demographic characteristics of the nationwide population of college and high school students. As such, the survey dataset may be biased in unknown ways and researchers should know they cannot draw conclusions about the larger population of recent graduates based on the survey dataset. PIL bears no responsibility for the interpretations or conclusions reached or presented based on third-party analysis of the PIL data.

8.3 Limitations of Liability

The study dataset is provided "as is" without any warranty of any kind, either express or implied, arising by law or otherwise, including but not limited to warranties of completeness, non-infringement, accuracy, merchantability, or fitness for a particular purpose. PIL expressly disclaims and shall have no liability for any errors, omissions, inaccuracies, or

interpretations in the survey dataset.

The user assumes all risk associated with the use of the data. The user agrees that in no event shall PIL be liable to the user or any third party for any direct, indirect, special, incidental, punitive or consequential damages including, but not limited to, damages for the inability to use equipment or access data, loss of business, loss of revenue or profits, business interruptions, loss of information or data, or other financial loss, arising out of the use of, or inability to use, the data based on a theory or liability including, but not limited to breach of contract, breach of warranty, tort (including negligence), or otherwise, even if the user has been advised of the possibility of such damages.

9. About the Codebook

The 14-page codebook provides researchers using the dataset with details about the 17 questions in the public dataset, including the full text of each question, response categories, frequency counts, variable names and missing data calculations. Data broken out by institution (Question 13) are not available, per IRB requirements, for this study. Questions collecting personal identifiers – Twitter screen names (Question 19), names and email addresses for a follow-up interview (Question 20), or entry into a drawing for an Amazon gift card (no question number assigned) – were eliminated from the public dataset before data analysis to protect the privacy of study participants.

Summary statistics (minimum, maximum, mean, median, mode, and standard deviation) are provided for the majority of questions, but are not available for non-numeric variables. Nominal and ordinal variables are only described with minimum and maximum, and character variables have no summary statistics at all. A listing of frequencies in table format may not be given for every variable in the codebook. Character variables do not have frequencies provided. However, all variables in the dataset are present and sufficient information about each variable is given.