Numeracy
Advancing Education in Quantitative Literacy

Volume 15 | Issue 1 Article 2

2022

Surveying the Landscape of Numbers in U.S. News

John Voiklis
Knolohy, johnv@knology.org
Jena Barchas-Lichtenstein
Knology, jenabl@knology.org
Bennett Attaway
Knology, bennetta@knology.org
Uduak G. Thomas
Knology, uduakt@knology.org
Shivani Ishwar
Knology

See next page for additional authors

Surveying the Landscape of Numbers in U.S. News

Abstract

The news arguably serves to inform the quantitative reasoning (QR) of news audiences. Before one can contemplate how well the news serves this function, we first need to determine how much QR typical news stories require from readers. This paper assesses the amount of quantitative content present in a wide array of media sources, and the types of QR required for audiences to make sense of the information presented. We build a corpus of 230 US news reports across four topic areas (health, science, economy, and politics) in February 2020. After classifying reports for QR required at both the conceptual and phrase levels, we find that the news stories in our sample can largely be classified along a single dimension: The amount of quantitative information they contain. There were two main types of quantitative clauses: those reporting on magnitude and those reporting on comparisons. While economy and health reporting required significantly more QR than science or politics reporting, we could not reliably differentiate the topic area based on story-level requirements for quantitative knowledge and clause-level quantitative content. Instead, we find three reliable clusters of stories based on the amounts and types of quantitative information in the news stories.

Keywords

quantitative reasoning, quantitative literacy, journalism

Cover Page Footnote

John Voiklis is a cognitive and social psychologist who leads behaviors research at Knology. Jena Barchas-Lichtenstein is a linguistic anthropologist who leads media research at Knology and is co-PI of Meaningful Math. Bennett Attaway is a researcher at Knology who focuses on explaining complex concepts to general audiences. Uduak Grace Thomas is Communications Manager at Knology. Shivani Ishwar was a researcher at Knology and is currently a Digital Analytics Fellow at United Nations Global Pulse. Patti Parson is Managing Producer of the *PBS NewsHour* and PI of Meaningful Math. Laura Santhanam is data producer at *PBS NewsHour*. Isabella Isaacs-Thomas is a news assistant at *PBS NewsHour*.

Authors

John Voiklis, Jena Barchas-Lichtenstein, Bennett Attaway, Uduak G. Thomas, Shivani Ishwar, Patti Parson, Laura Santhanam, and Isabella Isaacs-Thomas

Introduction

People have been counting for at least 50,000 years; that is more or less how far back the archeological evidence goes (e.g., Schmandt-Besserat 1992). Along the way, people have created professions and government agencies devoted to quantifying almost every human activity. News providers feel compelled to report many of these quantities.

Earlier research has found consistently that a large proportion of news stories include at least some quantitative or statistical information. For example, in an audit of one major newspaper, Maier (2002) found that quantitative reasoning is required by about half of all news stories. A study of British television, radio, and online news found that 22% of stories contained references to statistics, specifically (Cushion et al. 2017). And in an experimental study of news interpretation, Koetsenruijter (2011) found that using numbers rather than descriptive quantifiers like "many" makes people somewhat more likely to judge a story as credible, as does adding more numbers. This effect may be due to widespread (albeit false) assumptions that quantification is inherently objective (cf. Porter 1995; McConway 2016).

Both journalists and audiences are prone to misinterpreting numbers. Maier (2002) identifies 11 types of common errors in numerical content, including errors of computation and errors of interpretation. Utts (2003) identifies seven common statistical misconceptions which affect both journalists in their presentation of data and the public in drawing inferences from it. Gal (2002), rather than focusing on specific errors readers make, outlines five "statistical literacy" skills necessary to interpret and form conclusions from statistical information in news content. Specifically, adults should (1) know why data are needed and how data can be produced; (2) be familiar with basic terms and ideas related to descriptive statistics; (3) be familiar with basic terms and ideas related to graphical and tabular displays; (4) understand basic notions of probability; and (5) know how statistical conclusions or inferences are reached. To what extent are these expectations for quantitative knowledge and skills evident in the reporting of the news?

We are a team of journalists (PBS NewsHour) and social scientists (Knology) engaged in a long-term participatory collaboration (Barchas-Lichtenstein et al. 2020) that aims to understand how U.S. adults' news consumption impacts their quantitative reasoning (Barchas-Lichtenstein et al. 2021). Quantitative reasoning is a practical skill set that involves making sense of numbers in context and using them to inform decisions (Karaali et al. 2016). We elaborate further on definitions in an agenda-setting piece published in this journal (Barchas-Lichtenstein et al. 2021).

As part of this larger project, we sought to gather a sense of the current news landscape, focusing on four wide topic areas—economic, political, science, and

health news—where data-based reporting is common. Our goal was to assess the amount of quantitative content present in a wide array of media sources, and the types of quantitative reasoning required for audiences to make sense of the information presented. Eventually, this will enable us to compare adults' quantitative literacy with the demands placed on them by the news they consume, as well as to identify where and how news producers may be able to support their audiences' understanding.

The Present Research

Through the present study, we hoped to gain insight into how often "typical" news stories in four topic areas require quantitative reasoning from the reader, and in what ways. (In the "Data Collection" section of Methods, we elaborate on how we operationalized "typical" in our data collection. Put simply, we used a news aggregator to compile stories from a broad range of U.S. outlets.) We assessed the quantitative reasoning skills required for understanding individual clauses in a story and for making sense of the story as a whole. Specifically, we hoped to quantify the demands placed on readers of these stories, look for relationships between the story-level and clause-level quantitative reasoning requirements, and compare the types and extent of reasoning required across different topic areas. We asked the following descriptive research questions:

RQ1: How much quantitative reasoning (as operationalized in the "Coding and Codebook" section) do "typical" news stories require from readers?

RQ1.1: Are there differences in the type of quantitative reasoning required in different meta-data categories—topic areas, producing source type (legacy media outlet or an online-first publication), and medium?

RQ2: What relationships, if any, exist between quantitative reasoning at the conceptual/story level and at the clause level?

RQ2.1: How reliably do any such relationships organize news stories?

Methods

To answer these research questions, we followed a multistep process. First, we identified focal topic areas and examined what existing research says about their importance to life-long quantitative reasoning (Focal Topic Areas section). Then, we developed and executed a data collection strategy that would yield a representative sample of news sources across news-delivery platforms (print, television, online, etc.) (Data Collection section). Concurrently, we developed and applied a classification scheme (a "codebook" of classification "codes") for two nested units of analysis: news stories and their constituent clauses (Coding and Codebook section). Finally, in the Analyses of Codes section, we provide an overview of our analytic procedures.

Focal Topic Areas

The focal topic areas for this research were selected in conversations between journalists and social scientists on our team. Together, we identified four topic areas as particularly heavy in quantitative content: economics, science, health, and politics. The first three areas are more consistent in this regard; political news has considerably more quantitative content in presidential election years, because polling becomes a heavy focus of political news in those years. Because we were conducting this research in one such year, we found it important to include this topic area. These also largely map onto the topic areas where Cushion et al. (2017) found more statistical content, although Cushion et al. segmented the news stories they examined into narrower subject areas (for example, they considered "Energy" and "Environment" separately from "Science/technology"). Here, we map out some of the quantitative considerations central to reporting in each of the four areas.

Economic reporting focuses heavily on official statistics, and typically presents these statistics as self-explanatory, divorced from any mechanisms that cause change (Jensen 1987). That is, the economy is "consistently described as a set of variables. Economic actors, such as big corporations, small firms, wage earners, or consumers, are absent" (Jensen 1987, 19). Economists have also long noted that official economic statistics do not account satisfactorily for sampling error and various kinds of non-sampling error, leading to an illusion of more certainty than is warranted (Manski 2015). Journalists reporting on these official publications may have no information about uncertainty and thus may take these estimates as fact, reporting on fluctuations that may turn out to be insignificant.[1] And indeed, both Hope (2011) and Kleinnijenhuis et al. (2013) find that financial reporting can magnify uncertainty about the stock market, creating negative effects.[2] Similarly, Soroka et al. (2015) find that media shapes public opinion about the economy, and particularly that media reflects future change above all. At a more micro level, Gao and Corter (2020) find that audiences have better comprehension when change over time is presented in chronological order, yet economic journalists regularly report new values before the baselines. Here is one of many such examples from our data: "Consensus economists expect headline PCE will have risen 1.8% over last year in January, picking up from December's 1.6% year on year pace" (McCormick 2020).[3]

Meanwhile, Figdor (2017) argues that **science** journalists share responsibility with scientists for communicating uncertainty to the public. Figdor worries that

[1] See Irwin (2020) for new sources of uncertainty in economic estimates due to COVID-19.
[2] We are indebted to Nguyen and Lugo-Ocando (2016) for drawing our attention to this research.
[3] For ease, we've included references to all examples from our data set in a separate section within our references.

"epistemic failures" in science—behaviors like *p*-hacking[4] and adaptive sampling—are met with "epistemic vulnerability" in journalism: journalists may lack the expertise to assess the work done and thus be overly trusting, reporting on results obtained without integrity. Similarly, other research has found that journalists do not behave as authorities capable of choosing between competing truth claims, and that scientists are incentivized to overstate their claims (Lehmkuhl and Peters 2016). As journalists increasingly report on science in progress, including pre-prints of studies that have not yet been peer reviewed, they may struggle to identify which evidence is reliable (Dunwoody et al. 2018).[5] Recently, scholars have identified the promise of "weight-of-evidence" reporting strategies—which identify where the consensus of experts lies, "allow[ing] the journalist to present the array of truth claims in a way that acknowledges their presence but also makes clear what the bulk of experts believe to be true" (Kohl et al. 2016, 979).

Health journalists see statistics as one of the most important components of their stories, with only speaking to medical experts and defining technical jargon deemed more important (Hinnant and Len-Ríos 2009). By focusing on statistics, Hinnant and Len-Ríos conclude, journalists are leaving low-numeracy news users behind.[6] It is important to note that we collected our data in February 2020, at the time when COVID-19 was first receiving a lot of coverage in U.S. media.[7] Written pieces about COVID-19 have focused heavily on numbers while "gloss[ing] over the rather messy procedures used to create those numbers" (Best 2020, 4). That is, confirmed case and death counts were treated as objective, even though areas had different standards for attributing deaths to COVID and capacity to test and track cases. It has become increasingly clear in the months since we collected this data that the spread of COVID-19 cannot be understood without reference to economic and political decisions made at local, regional, and national levels (Briggs and Nichter 2009; Briggs 2011). Nor have journalists always succeeded in comparing apples to apples, sometimes reporting on absolute frequencies when population-adjusted proportions would be more apt (Ancker 2020). And health journalists also

[4] *P*-hacking involves reporting spurious results that happened to meet some threshold of statistical significance. As one journalist author notes, journalists may not even have heard this term, much less be able to identify the practice. A clear journalist-facing explainer is available at https://scienceinthenewsroom.org/resources/statistical-p-hacking-explained/

[5] While journalists have long had access to pre-prints, reporting on them has increased as a result of the COVID-19 pandemic (Fleerackers et al. 2021).

[6] Reyna et al. (2009) present a thorough review of the challenges of low numeracy and dense quantitative information in health.

[7] At that time, the U.S. public was not yet aware of widespread community transition at home, and U.S. media was still largely treating COVID-19 as a foreign problem with few if any domestic effects (cf. Benton and Dionne 2015; Benton 2016). Only seven states had announced at least one case by March 1, 2020; by March 17, 2020, all 50 states had confirmed cases. However, experts now believe there was already widespread community transmission at this point (Carey and Glanz 2020), and many states have since revised the dates of their first cases.

need to be mindful to put these numbers in context, including such information as the range of possible values, benchmarks or thresholds, and associated uncertainty (Ancker 2020).

As early as forty years ago, Crespi (1980) noted that **political** opinion polling has deep ties to journalism that has impacted both the strengths and weaknesses of their methods.[8] Other research has consistently found that journalists do not present sufficient methodological information for readers to make their own judgments about political polls (Bhatti and Pederson 2016; see Portilla 2016 for a review of many of these studies). Journalists are not sufficiently conservative when interpreting poll results, reporting on differences that are likely due to chance (Bhatti and Pederson 2016). Even where the law requires reporting on methodology, not all journalists comply—one study found that about one-third of Spanish news stories did not include legally required information (Portilla 2016). Portilla (2016) also found that newspapers may be somewhat more likely to report on methodology when they have commissioned the polls; Bhatti and Pederson (2016) found that journalists who included methodological information were no more or less conservative than others in their interpretation.

Data Collection

Selecting a representative sample of news sources across platforms is challenging. As of 2015, legacy news still accounted for about 2/3 of the top 25 news sites in the United States by unique visitor count (Pew Research Center 2015). These sites included ten newspapers, six broadcast television networks, one radio station site (NPR), and eight online-only sites, one of which was a pure aggregator. Current YouGov data (YouGov 2021) uses different metrics to rank sources, and their top 25 news sites still include 14 legacy outlets. Because it is difficult to assess the popularity of news sources across platforms, and because news aggregators have become an important source of news (Lee and Chyi 2015), we relied on a news aggregator, specifically Google News, to collect our data set.

We recognize both strengths and limitations of this choice. Some research (e.g., Weaver and Bimber 2008) found that Google News is more complete than major databases, and two recent studies (Haim et al. 2018; Nechushtai and Lewis 2019) found that there is relatively little implicit personalization in Google News. However, both studies also found that Google News has a high concentration of a small number of sources, particularly relative to their distribution. Similarly, a study of Apple News (Bandy and Diakopoulos 2020) found that human aggregation was more effective than algorithmic aggregation in selecting diverse sources and in focusing on "hard news" topics. In other words, these strengths and weaknesses

[8] In the same special issue on polls and the news media, Paletz et al. (1980, 499) observe, "The press seems obsessed with presidential elections, willing to publish polls on the subject no matter how irrelevant and inane."

may not be specific to Google News but rather typical of algorithmic aggregators in general.

Over a seven-day period that did not include any major U.S. holidays, with data collection happening at the same time each day, we scraped the content of the first 20 stories appearing in Google News in each of the following categories—Economy (referred to on Google News as "Business"), Science, Health, and Politics. We selected the first six for each day that were (a) from U.S. sources, (b) not opinion pieces, and (c) not duplicates of stories already in the data set. In addition, during the same period and at the same time, we selected the first three stories appearing in the same four categories on the PBS NewsHour website. In this way, we amassed a NewsHour corpus large enough to provide base-rates of quantitative content for NewsHour-related research activities without overly biasing our data set towards the NewsHour's linguistic choices.[9] The Supplemental Materials list all downloaded stories, authors, outlets, and reasons for inclusion or exclusion in the database (all supplemental documents are available for download at https://bit.ly/3jIF3K0).

To speed human coding, we automatically parsed each news report into clauses, defined operationally by the presence of periods, colons, semi-colons, exclamation points, or question marks. While linguistic analysis would normally use true clauses—grammatical units that contain a predicate and an overt or non-overt subject—as a unit of analysis, this operational definition allows us to use the same unit of analysis for human and machine coding. See the Supplemental Classification Protocol for more information about this process and associated challenges (again, https://bit.ly/3jIF3K0). As a final step, we located the articles on the sites where they were originally posted and downloaded them in PDF format to verify the presence or absence of graphics.

Coding and Codebook

We coded two nested units of analysis: stories and clauses. Researchers coded all stories in qualitative data analysis (QDA) software. We used Dedoose, which allows simultaneous access to the same data set. (See Hart and Achterman (2017) for a comparison between several commonly used QDA packages.)

Story-Level Codes. One set of codes takes the single news report as the unit of analysis. Arguably, these story-level codes represent the gestalt of the news report that guide story comprehension by constraining, and therefore facilitating, inferences about specific parts of the report (e.g., St. John 1992). This set of codes (Table 1) is based on the five key components of statistical literacy proposed by Gal (2002, 10).

[9] Because our goals include training a machine-coding algorithm, this balance was of particular concern.

Table 1
Story-Level Codes
To make sense of this story, does the reader need to . . .

Code	Description & Instructions
. . . know why data are needed and how data can be produced?	Answer yes if the story includes references to studies, research, and the collection of data. This category encompasses stories that reference research design and data analysis without being explicit as to why they use certain methods or quantified values.
. . . be familiar with basic terms and ideas related to descriptive statistics?	Answer yes if the story includes references to quantified values based on existing data, including comparisons between values, central tendencies and exceptions, and proportions and percentages. This category references stories that use existing data to describe current phenomena and realities, rather than predictions or likelihoods.
. . . be familiar with graphic and tabular displays and their interpretation?	Answer yes if the story includes visualizations of quantified values, including charts, graphs, or more complex visual displays. This category includes any story that includes one or more such visualization.
. . . understand basic notions of probability?	Answer yes if the story includes references to predictions, projections, or probabilities. These should be quantifiable, rather than pure conjecture about possibilities for the future. This category includes stories about polling, weather forecasts, and economic projections.
. . . know how statistical conclusions or inferences are reached?	Answer yes if the story provides references to inferences or conclusions made based on quantified values, without explicit explanation of how those conclusions were reached. This category is often linked to predictive or descriptive statistics.

In preliminary coding rounds, raters had near-total agreement on all story level codes (there were only two instances of disagreement over whether a story should receive a certain code, and agreement was quickly reached through discussion). As a result, we did not formally calculate inter-rater reliability for these codes.

Clause-Level Codes. A second set of codes takes the clause as the unit of analysis. This series of codes was developed bottom-up by the authors (several of whom have statistical and/or computational training) through discussion and iteration. The process was inductive and abductive, based on close reading of a dozen NewsHour stories (three from each topic area). After initial development of the codebook, we sought review and feedback from external researchers (project evaluators Jim Hammerman and Eric Hochberg and project advisors James Corter, Danny Bernard Martin, and Darryl Yong). Table 2 presents an abridged version of the codebook; the full codebook is available in Appendix A.

After all stories were coded, researchers decided to merge two of these codes—"Sampling, Representativeness, and Generalizability" and "Enumeration"—into a single overarching Research Methods code.

Table 2
Clause-Level Codes

Code	Description
Official Statistics and Official Statistics Organizations	Reference to official statistics (cf. Gal & Ograjenšek 2017) as well as any organization or agency whose mandate includes the publication of official statistics, including governmental and non-governmental polling organizations
Comparison	Comparison of statistical quantities (including proportions, means, etc.) across populations or topics—refers to comparisons of one value to a different value; includes comparisons to some norm or expected value, even where the base rate is left unspecified; must refer to values that are quantifiable and quantified, even if the specific quantities are not made explicit
Proportion or Percentage	References to proportions that may or may not include explicit percentages; also includes unquantified references to rates when these rates are clearly designated within the text as quantifiable
Central Tendencies and Exceptions	Includes references to averages—either means or medians—whether the type of average is explicit or implicit; includes references to modes; must refer to values that are quantifiable and quantified, even if the specific quantities are not made explicit; also includes outliers and exceptions from the norm, assuming that they imply a typicality or average even if not directly stated
Variability, Concentration, and Variation	Reference to a concentration or uneven distribution of a statistical phenomenon; differs from sampling and representativeness because variability and concentration are challenges for sampling and representativeness, but are not themselves used to generalize to a larger population
Risk and Probability	References to risk, likelihood, or probability (e.g., of exposure to danger, harm, or loss; or future events or outcomes); includes predictions and forecasts in scientific or political senses; must be about quantifiable statistical values, not the concept of probability in a more general sense
Magnitude and Scale	References to scale, amount, or number of values being examined or assessed (cf. Yarnall and Ranney 2017); includes raw numbers and sometimes approximate numbers; small numbers are also "magnitude" if they're specific
Sampling, Representativeness, and Generalizability	Reference to research methods that use a sample population to generalize across a larger population, whether explicit or implicit; note that "representativeness" is distinct from diversity or "representation," though the two can overlap in certain contexts
Enumeration (and Inclusion/Exclusion Criteria)	Refers to research methods that use full enumeration, i.e., counting, rather than estimates derived through statistical inference; only use this code if you are certain that this is the method by which the number was derived; specific criteria for what IS or IS NOT counted goes under this code

After the full data set (230 stories) was collected, we followed the most recent guidelines for reliable content coding (O'Connor and Joffe 2020). Each of three pairs of Knology researchers (i.e., each possible pairing of authors JBL, EA, SI) was assigned between 24 and 26 stories at random, so that inter-rater reliability could be calculated for a set of 78 stories, one-third of the full data set. The sample was stratified such that NewsHour and Google News stories were assigned proportionately. We calculated Gwet's AC 1 for each pair of coders, which reached

substantial levels (≥0.8) for all but one code (0.7≤AC1≤0.8). One of the coders (author JBL) reconciled all disagreements resulting from the inter-rater reliability, and a second coder (author EA) coded all remaining stories in the database. One fully coded story is reproduced in Appendix B.

Analyses of Codes

The analysis of the codes unfolds over several steps. After summarizing the general characteristics of the news-story corpus, we provide descriptive statistics to summarize the two sets of codes as a general description of our news story corpus--overall (RQ1) and by meta-data categories—topic areas, producing source type, publication medium (RQ1.1). Next, we show that each set of codes (at the story level and at the clause level) can be reduced to a smaller number of summary variables—specifically, principal components—that capture bundles of intercorrelated concepts/codes (RQ2), then show that any patterns in the principal component scores are negligibly related to meta-data categories—topic areas, producing source type, medium. Instead, we show that patterns in the principal component scores organize groups of articles that were reported with similar expectations for story-level requirements of quantitative knowledge and similar clause-level quantitative content. Finally, we characterize these groups of stories based on the patterns in their quantitative content and knowledge expectations.

The details of the statistical techniques used for grouping codes and grouping stories are provided within each results section for ease of reference.

Results

Characteristics of the Data Set

Before turning to the analysis of classification codes, we outline the characteristics of the data set we collected. Our data set contains a total of 230 stories collected between February 18 and February 24, 2020. During this period, a single topic—the spread of COVID-19 and associated economic shocks—was already somewhat dominant; over one-fourth of the stories address this topic. The impending pandemic likely affected how news stories were reported, perhaps increasing the focus on case counts and projections and, thus, skewing the results.

Of 230 stories, 167 were collected through Google News while 63 came from PBS NewsHour, including syndicated stories from the Associated Press. Only 40 stories were videos, while the other 190 were in text format. They were well-balanced among the four topic areas, ranging from 53 science stories to 61 politics stories.[10] However, the politics stories were somewhat longer than stories in the

[10] It is important to note that we removed duplicate stories from the data set. Some stories appeared on multiple dates, and some stories appeared in multiple topic areas. Some of the difficulty in

other topic areas (containing a median of 45 clauses, compared to 27–31 for other topics).

The 230 stories in the data set were produced by 74 distinct outlets. Two researchers classified each outlet into two categories: we noted whether the outlet was a legacy media source (i.e., print or broadcast) or an online-first publication in case there were differences between them. The resulting Producing Outlet Type meta-data variable included 38 outlets that were classified as online-first, 35 as legacy media, and one that did not fit into either category.

Each of the four topic areas shared some general characteristics that the coders noted during the classification process:

Economy articles were generally fairly code dense, reporting on such topics as interest rates, profits, and stock prices. Many articles addressed change over time in these economic variables, often due to some external change. Most articles also included some sort of quantifiable prediction or forecast. The coder noted that they did not code references to "the Dow" as Central Tendencies unless it was explicitly specified as "the Dow Jones Industrial Average" because readers unfamiliar with this number might not know how it is calculated.

Health stories were overwhelmingly about COVID-19 or the seasonal flu. These stories often included references to case counts "confirmed" by local health authorities, sometimes including considerable detail about who was or was not included in such counts. We also saw a number of references to clusters and outbreaks in specific areas. Some of these stories also focused on the likelihood of WHO declaring COVID-19 a pandemic (which it eventually did). Comparisons between the two diseases (COVID-19 and influenza) were also somewhat frequent.

The numbers that showed up most frequently in **politics** stories typically referred to years or dollars. Compared to Economy and Health stories, magnitudes were less precise: "thousands of voters" or "millions of dollars." Even politics stories that had a few codes present did not necessarily require any statistical literacy to be understood. Stories that did require statistical literacy were typically about elections, addressing issues like demographic differences in political attitudes, "electability," and various candidates' likelihood of winning.

Science stories about research studies had quite a few statistical concepts. However, space exploration was a big topic the week we collected data due to announcements that SpaceX was planning to launch tourists into orbit and that Japan was planning a Phobos mission. Stories about this announcement typically only included information about distances, costs, and perhaps comparisons to earlier missions.

distinguishing between topic areas that we discuss later is largely due to the lack of discrete boundaries between topic areas.

How Much Quantitative Reasoning Do "Typical" News Stories Require from Readers? (RQ1)

With these characteristics in mind, we now turn to the analysis of the classification codes.

Figure 1 summarizes the corpus in terms of story-level knowledge, showing the proportions of stories that received each story-level code. Differences between code occurrences are obvious from a simple visual analysis. Almost two-thirds of the stories required knowledge of why data are needed and how they are produced, as well as knowledge of descriptive statistics. Fewer than half of the stories required familiarity with probability and inferential statistics. Only about one tenth of the stories required familiarity with data visualization.

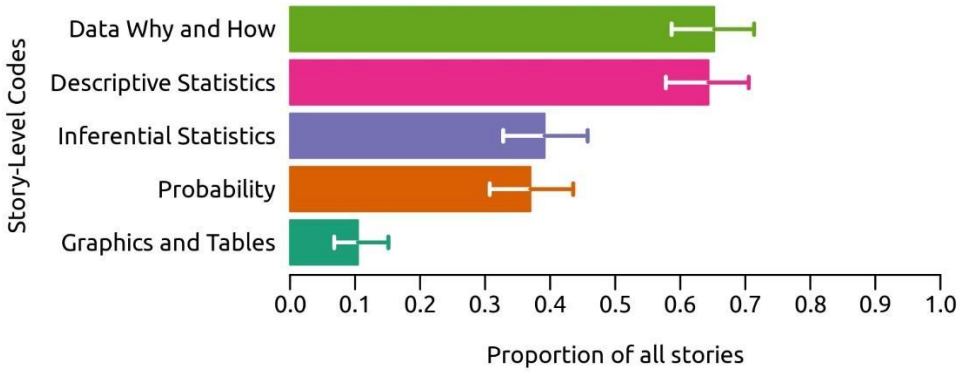

Figure 1. Summarizing the corpus in terms of story-level knowledge requirements: the proportions of stories assigned each story-level code. Whiskers indicate the 95% confidence interval for estimated proportion.

Figure 2 summarizes the corpus in terms of clause-level quantitative content. Panel (A) shows the proportions of stories where a clause-level code occurred at least once. Panel (B) shows the average (median) numbers of clauses per story that received a particular clause-level code.[11] Again, visual inspection suffices to discern differences between code occurrences. The two most common codes, Magnitude and Comparison, appeared in nearly every story. On average, stories contained between 5 and 7 clauses coded as Comparison and 4 or 5 clauses coded as Magnitude. This difference may be an artifact of segmenting stories into clauses: comparisons sometimes require multiple clauses to express a single comparison (e.g., *P*% of sample *S* plan to do activity *A*. However, only *Q*% follow through with

[11] Specifically, Figure 2, Panel (B) shows the 95% confidence interval for estimated median number of clauses per story receiving a particular clause-level code. In all cases, except for the Comparison code ($\tilde{x}$=6), the median values fell either at the upper or lower limit of the Confidence Interval.

their plan). The likely objects of those comparisons—Proportion, Variability, and Risk—were the next most common codes: each appearing in approximately two-thirds of the stories in the corpus. On average, stories included one or two clauses that received these codes. Meanwhile, Research Methods, Central Tendencies, and Official Statistics appeared in only about half of all stories, with an average of one or two clauses receiving these codes.

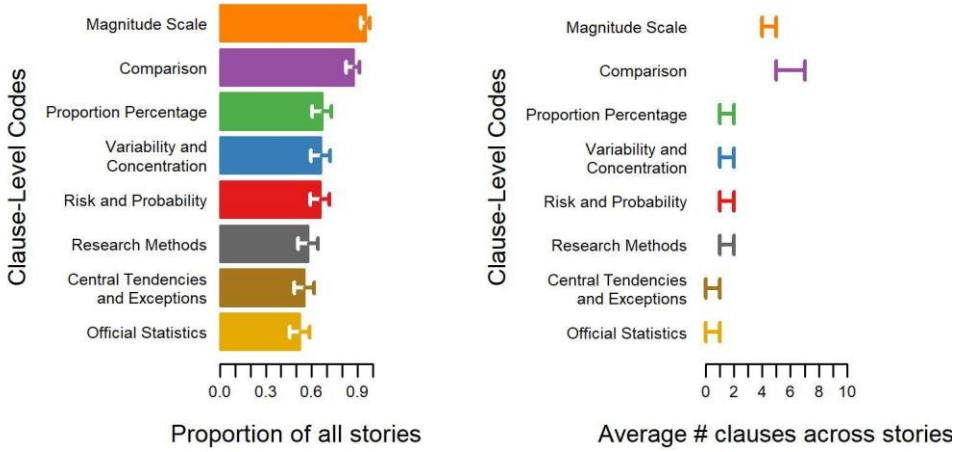

Figure 2. Summarizing the corpus in terms of sentence-level quantitative content. Panel (A) shows the proportions of stories where a sentence-level code occurred at least once. Whiskers indicate the 95% confidence interval for estimated proportions. Panel (B) shows the 95% confidence interval for estimated median number of clauses per story that received a particular clause-level code.

Are There Differences in the Type of Quantitative Reasoning Required in Different Meta-data Categories—Topic Areas, Producing Source Type, and Medium? (RQ1.1). Figure 3 organizes the story-level codes by three meta-data categories: topic areas, producing source type, and medium. Panel (A) shows the proportion of stories in each of the topic areas that received each of the story-level codes. Economy and health news required the most quantitative knowledge to interpret the stories. Almost all economy and health stories required knowledge of why data are needed and how they are produced, as well as of descriptive statistics. Only about half of science stories, and fewer than half of politics stories, required this knowledge. More than half of economy and health stories required familiarity with probability and inferential statistics, while one-third or less of science and politics stories did. Approximately one quarter of stories about the economy required familiarity with data visualization, that is, graphs and tables. Only about one tenth of the stories in the other three topic areas made such demands on audiences. Panel (B) shows the proportion of stories in each of two types of producing outlets (legacy outlets, with a print or broadcast presence, and online-

first outlets) that received each of the story-level codes. A simple visual analysis shows that online-first outlets required somewhat more quantitative knowledge across all five codes. And Panel (C) shows the difference between text and video stories. Visual inspection suggests that all differences between media might be attributable to chance occurrence, as apparent from the overlapping 95% Confidence Intervals for the estimated proportions.

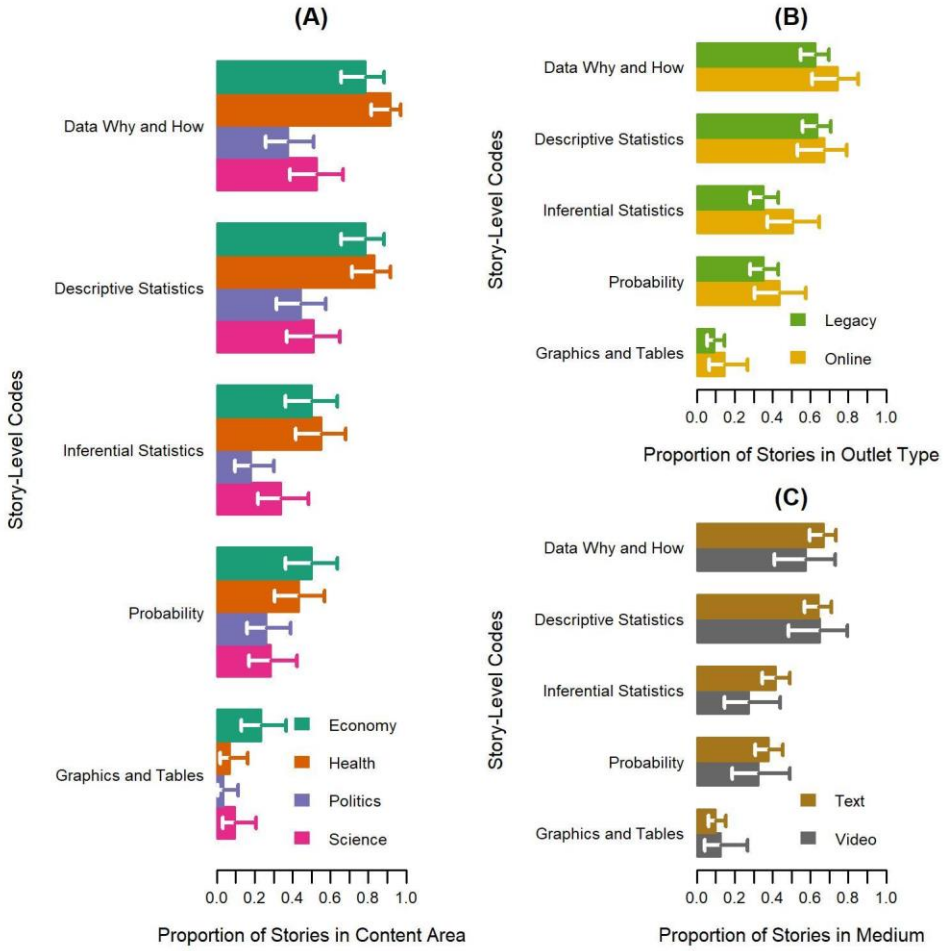

Figure 3. Distributions of story-level code proportions across meta-data categories—topic areas (Panel A), producing source (Panel B), medium (Panel C). Whiskers indicate the 95% confidence interval for estimated proportions.

Turning to clause-level codes, Figure 4 organizes the story-level codes by three meta-data categories: topic areas, producing source type, and medium. Panel (A) shows the proportions of stories in each of the topic areas where a clause-level code occurred at least once. As with corpus-level summary, Magnitude appeared in

nearly every story across all topic areas. Comparison appeared in nearly every economy and health story, and approximately three-quarters of politics and science stories. The occurrence of the remaining codes decreased stepwise as observed in the corpus-level summary. There was also considerable variation between topic areas, but in general, codes appeared in a greater proportion of economy and health stories than in science and politics stories. Panels (B) and (C) show the proportions of stories from each producing source and in each medium, respectively, where a clause-level code occurred at least once. As apparent, there was relatively little variation between legacy and online-first outlets, or between text and video.

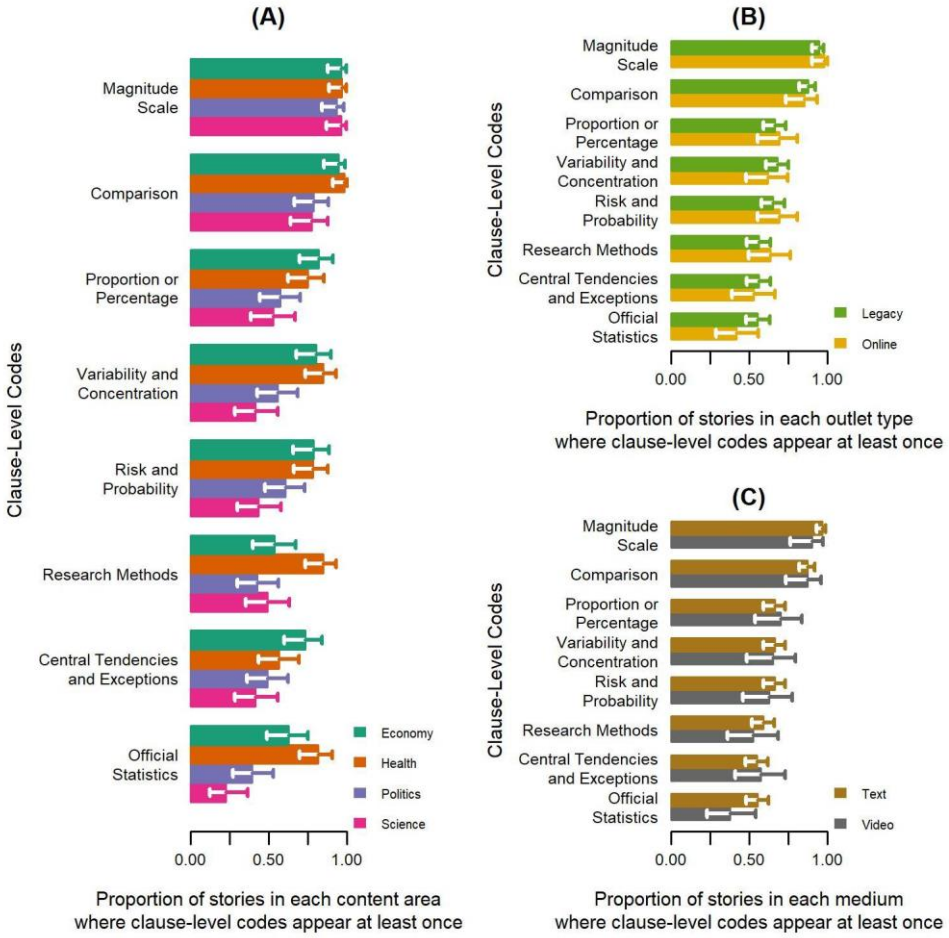

Figure 4. Distributions of clause-level code proportions across meta-data categories—topic areas (Panel A), producing source (Panel B), and medium (Panel C). The bars show the proportions of stories in each meta-data category where the clause-level codes appear at least once. Whiskers indicate the 95% confidence interval for estimated proportions.

We also looked at the distributions of clause-level code frequencies (average number of clauses per story that received a particular clause-level code) across the meta-data categories. A figure is available in the Supplemental Results document. While the distribution of frequencies differs in some of the minor details, it does not change the overall patterns observed in Figure 4: a stepwise decrease in the frequency of codes, with a similar pattern of differences between topic areas, producing source, and medium.

What Relationships, If Any, Exist Between Quantitative Reasoning at the Conceptual/Story Level and at the Clause Level? (RQ2)

While individual codes encapsulate separable aspects of quantitative reasoning, those aspects are often dependent on one another. Therefore, one would expect codes to co-occur with one another. In fact, among the 162 stories that received at least one story-level code, 145 stories (90%) received two or more codes. Similarly, among the 4,249 clauses that received at least one clause-level code, 2,088 clauses (49%) received two or more codes. More than half of the clauses in our data set received no clause-level codes. See Figure 5 for the distribution of story-level codes and stories and clause-level codes and clauses.

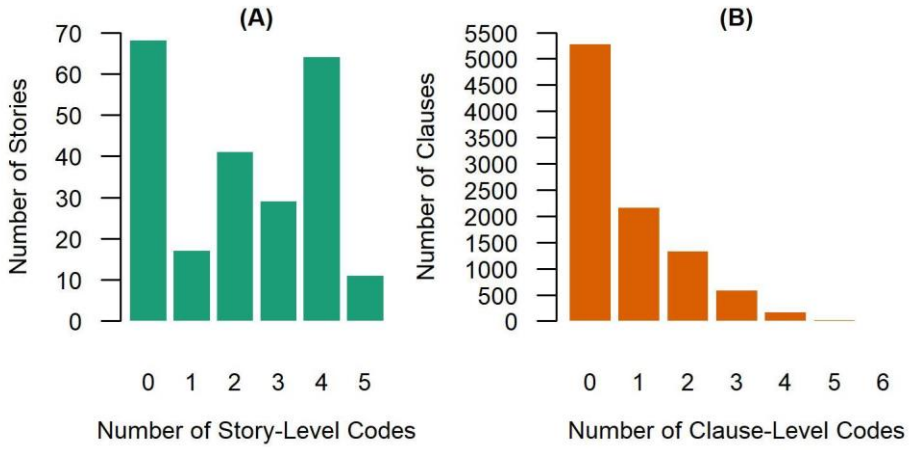

Figure 5. Number of story-level codes assigned to story and the number of clause-level codes assigned per clause in the news story corpus.

More importantly, codes consistently co-occurred. We used two measures of consistency: Cronbach's α statistic and the average (polychoric[12]) correlations between codes at the story level and at the clause level, respectively. For story-level codes, the 95% Confidence Interval of Cronbach's α was $0.78 \leq \alpha \leq 0.84$, with an average correlation of $0.55 \leq \bar{r} \leq 0.82$. This level of consistency suggests that one might summarize the story-level codes with a single summary variable. We used parallel analysis (Zwick and Velicer 1986) to compare the story-level data to simulated data and estimate the optimal number of principal components (summary variables) to extract with principal component analysis (PCA). When comparing the actual data to the simulated data in the scree plot (Fig. 6, Panel (A)) only the first component has an eigenvalue greater than 1 and greater than the threshold based on the simulated data. In fact, a single-component PCA solution accounted for a large proportion of the variance in the data (79%).

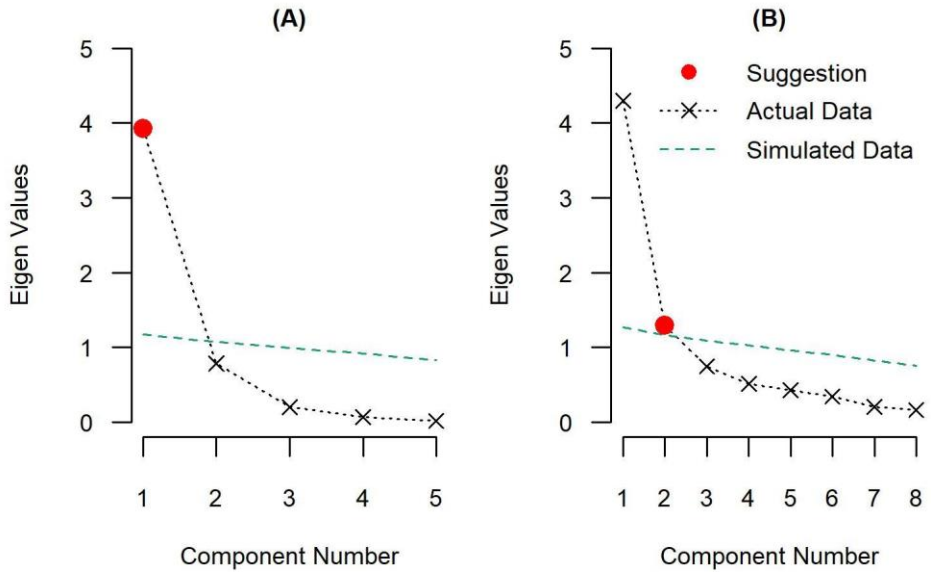

Figure 6. "Scree" plot of eigenvalues for each component extracted from the data. Panel (A) shows the estimates for story-level codes. Panel (B) shows the estimates for clause-level codes.

One can read the "loadings" shown in Figure 7, Panel (A) much like correlations between the individual codes and the principal component. All codes are strongly associated with the component (the lowest loading is 0.62), and all knowledge requirements, except for familiarity with graphical and tabular displays, are almost entirely redundant with the component (loadings greater than 0.9). In

[12] Polychoric correlation is used with dichotomous and ordinal data to estimate what the Pearson correlation would be if variables were on a continuous scale.

other words, one can summarize the story-level data simply as the amount of quantitative knowledge needed for story comprehension.

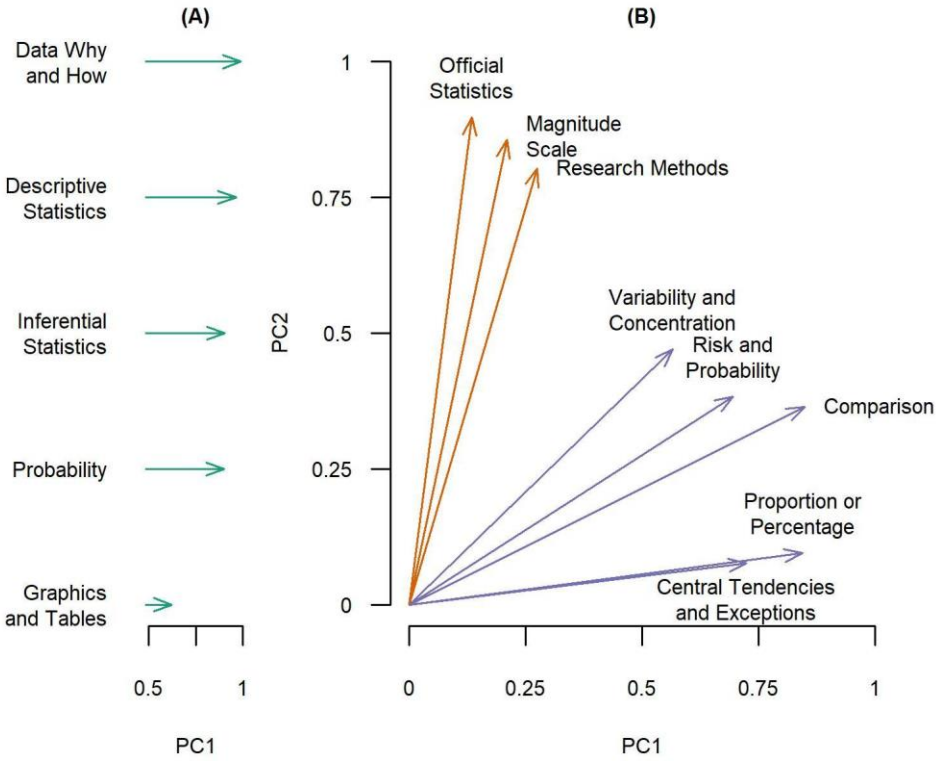

Figure 7. The strength of the relationship (loadings) between the classification codes and the underlying summary variables (Principal Components).

Clause-level codes also consistently co-occurred. The 95% Confidence Interval of Cronbach's α was $0.81 \leq \alpha \leq 0.87$, with an average correlation of $0.31 \leq \bar{r} \leq 0.59$. Nevertheless, the lower average correlation suggests that one might need two or more principal components to summarize the clause-level codes.

A parallel analysis suggested two components (see Fig. 6, Panel (B)). A two-component PCA solution accounts for 70% the variance in the data (see Fig. 7, Panel (B)). One component (PC1) combines the co-occurrences between five codes: Comparison, Proportion or Percentage; Central Tendencies and Exceptions; Variability, Concentration; and Variation, Risk, and Probability. A second component combines the co-occurrences between the remaining three codes: Official statistics and official statistics organizations; Magnitude and Scale; and Research Methods. For the sake of brevity, we call the first component the **Comparison** dimension because it organizes the stories along a continuum of how much of the content refers to the quantities of comparison (including comparative

risks and probabilities). Similarly, we call the second component the **Magnitude** dimension because it organizes the stories by how much of the content refers to quantities of scale, amount, or number; the providers of these figures; and the procedures they used to measure these magnitudes.

How Reliably Do Any Such Relationships Organize News Stories? (RQ2.1). In the section on RQ1.1, we noted some rough differences between meta-data categories in the occurrence of story-level and clause-level codes. Differences were especially apparent between topic areas—for example, audiences need more quantitative knowledge to interpret economy and health news than to interpret politics or science news. Meanwhile, differences were small and often negligible when comparing producing outlet types and media types. Here, we used discriminant analysis models (DA; Tufféry 2011) to test whether the dimensions (knowledge requirements and comparison-related and magnitude-related content) that summarize the co-occurrences between codes might differentiate the meta-data categories. We relied on two statistics: (1) the correlation ratios (η^2) between the explanatory variables (the coordinates of the news stories on the dimensions, i.e., the principal component scores) and meta-data categories as a measure of the discriminant power of the explanatory variables[13]; and (2) the rate at which the DA models misclassify randomly-selected sub-samples of data (a model with an error rate<20% is considered reliable, akin to statistical power of 80%).

For topic areas, the discriminant power of the component scores was modest ($0.04 \leq \eta^2 \leq 0.12$) and the resulting model made exceedingly poor predictions (error rate=65%). For both producing outlet type and medium, the discriminant power of the component scores was negligible (all $\eta^2 \leq 0.01$) and the resulting model could not at all differentiate legacy from online outlets (all were classified as "legacy") or text from video (all were classified as "text"). All in all, we cannot reliably differentiate the meta-data categories—topic areas, producing outlet type, medium—for the news story corpus based on story-level requirements for quantitative knowledge and clause-level quantitative content.

The fact that meta-data does not differentiate quantitative news does not preclude using the three dimensions of code co-occurrence to organize and differentiate quantitative news. We used Latent Profile Analysis (LPA; Fraley and Raftery 2002) to discover and extract clusters of stories (among the 230 stories) that shared similar combinations of story-level and clause-level component scores. We used three criteria to select the optimal number of clusters: we looked for (1) the smallest number of clusters that would minimize both (2) Bayesian Information Criterion (BIC; a commonly used statistic for latent variable modeling; Vrieze

[13] The correlation ratio compares the statistical dispersion within categories against the dispersion in the sample or population; specifically, it is the ratio of the within-category standard deviations to the overall standard deviation in the data.

2012) and (3) the rate of misclassification using DA for external validation of the cluster solution.

As apparent in Figure 8, the BIC values exhibit three local minima for models of five, seven, and ten clusters. The five-cluster model has the lowest BIC value within the range of clusters that would reduce the likelihood of segmenting the news corpus into underpopulated groupings.

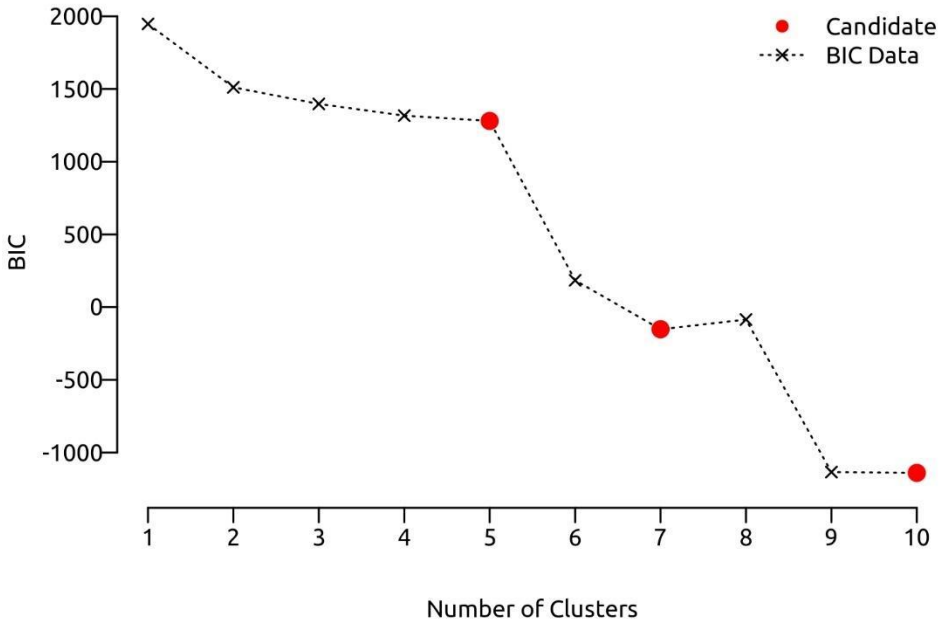

Figure 8. Plot of Bayesian Information Criterion values for "mixture models" with increasing numbers of clusters. Red points indicate candidate models for further testing.

We cross-validated the output of the models with five, seven, and ten clusters using discriminant analysis. We ran iterative tests of the models using 10,000 subsamples of the data, from which we calculated the 95% Confidence Interval on the mean error rate. The five-cluster model performed best on this criterion: error rate=19% for five clusters, error rate=21% for seven clusters, and error rate=27% for 10 clusters.[14] In fact, only the five-cluster model was cross-validated with an error rate below the 20% rule of thumb. That said, all models yielded some underpopulated clusters. Clusters with too few members may fail to yield reliable summary statistics. When characterizing the five-cluster solution, we will focus on

[14] We report single values because the upper and lower extremes of the 95% Confidence Intervals differed, at most, by 0.002.

the three clusters containing more than 30 news stories: clusters 1 (n=67), 3 (n=75), and 4 (n=74).

To begin characterizing the five-cluster model, we first examined the discriminant power of the component scores. All exceeded chance occurrence (p<0.001) and ranged from medium (η^2=0.49) for the Magnitude dimension, to substantial (η^2=0.70) for the Comparison dimension, to almost perfect (η^2=0.86) for the story-level knowledge requirement dimension. Figure 9 shows the distributions of the principal component scores for Clusters 1, 3, and 4.

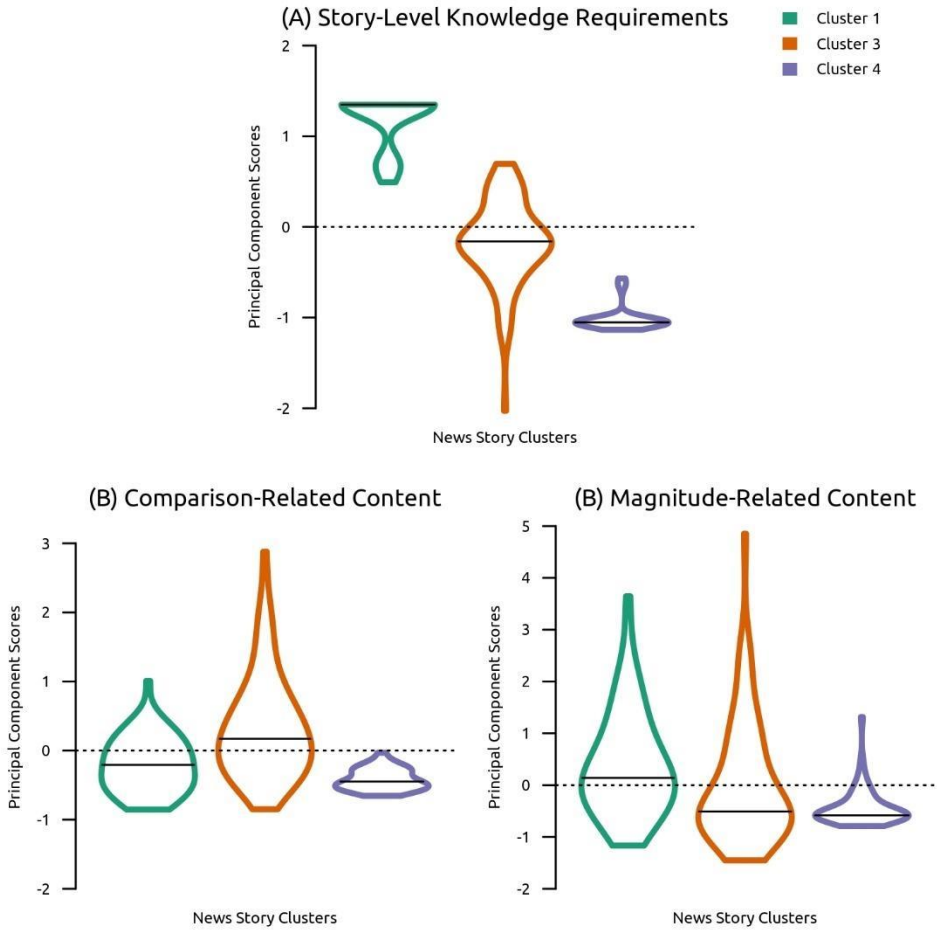

Figure 9. Violin plots of the three profiles. These plots show both a median (horizontal solid line) and the probability density of the data (shape of the "body").

The distributions in Figure 9 are shown as violin plots, the "bodies'' of which offer several features for visual analysis. The length shows the estimated range of the and the shape shows where to find the bulk of the observations (emphasized

with the horizontal line marking the median). For reference, imagine the violin plot for the normal distribution: it would look a bit like a lemon, with a pronounced belly at the center and short, symmetric protrusions at the top and bottom. A visual analysis corroborates the discriminant power analyses: the clusters exhibit distinct shapes on all dimensions, but clear separation in the story-level knowledge requirements. and the shape shows where to find the bulk of the observations (emphasized with the horizontal line marking the median). For reference, imagine the violin plot for the normal distribution: it would look a bit like a lemon, with a pronounced belly at the center and short, symmetric protrusions at the top and bottom. A visual analysis corroborates the discriminant power analyses: the clusters exhibit distinct shapes on all dimensions, but clear separation in the story-level knowledge requirements.

To understand what these statistical differences mean at the concrete level of the news stories themselves, it may help to examine the "prototypical" stories in each cluster. By "prototypical," we mean the news story in a cluster with the smallest Mahalanobis distance (Mahalanobis 1936) from the centroid values on the three principal components for the cluster. Table 3 displays the headline, topic area, and counts for both story-level and clause-level codes for the "prototypical" stories in the more reliable clusters.[15] The prototype of Cluster 1 addresses the complex economic factors associated with hydraulic fracturing for natural gas. The prototype of Cluster 3 represents an early attempt to quantify the spread of COVID-19 while the numbers remained small and seemingly easy to calculate and apprehend. The prototype of Cluster 4 reports a space exploration story with little reference to the many potential economic and engineering quantities involved. The code counts for each prototype clearly show a stepwise decrement in quantitative knowledge requirements and quantitative content.

Table 3
Prototypical Stories in Three Reliable Clusters

	Title	Content Area	*n* Story Codes	*prop.* Comparison-related Clauses	*prop.* Magnitude-related Clauses
Cluster 1	Fracking debate causes tremors in battleground Pennsylvania	Economy	4	0.50	0.56
Cluster 3	Italy's novel coronavirus cases rise to 17 as cluster emerges	Health	2	0.36	0.36
Cluster 4	Japan greenlights mission to bring back sample of Mars moon Phobos	Science	0	0.23	0.15

[15] We included topic area simply to disambiguate headlines, even though the principal components do not reliably differentiate topic areas.

Discussion, Implications, & Conclusions

When all is said and done, the news stories in our sample can largely be classified along one single dimension: the amount of quantitative information they contain. Simply put, stories with more quantitative content at the clause level also required more quantitative reasoning at the conceptual level (**RQ2**).

In particular, all story-level codes were highly correlated with one another. However, visualization occurred considerably less frequently than the other four. This result implies that journalists use graphs and tables differently than other quantitative information. Presumably, at least some journalists make use of visual ways of representing information specifically for non-technical audiences. Combining words and visualizations reduces cognitive effort and improves memory for the topic (for reviews, see Clark and Paivio 1991; Pearson and Kosslyn 2015), and visual presentation of information may also be more effective for audiences with less quantitative background (Attaway et al. 2020).

As a journalist author (LS) observes: when deciding and designing data visualizations, journalists are pushing against an array of constraints that can feel comical at times. In whole or in part, deadlines, staff, computer program updates, and machine and internet functionality can present a journalist with a sudden crisis that demands immediate resolution. To the degree possible, templates and processes can be developed to smooth out some unpredictability, enable greater efficiency, and yield more productive results. For example, in polling coverage with deadlines rapidly approaching before responses grow stale, visual journalists plug datapoints into pre-existing templates that have been thoughtfully constructed so as to clearly communicate the main idea behind choosing to illustrate those statistics in the first place. But the aforementioned constraints also demand a key function of journalism: to distill the most information to its most essential elements and share those findings with one's audience. That means choices need to be made quickly about what numbers matter most and must be shared with readers, viewers, listeners, or anyone who may encounter these stories once published or aired.

Meanwhile, quantitative clauses largely fell into one of two groups: those that included references to **magnitude** and associated concepts and those that included **comparisons**. Other codes typically correlated strongly with at least one of these two codes. A recent video by the U.S. Census Bureau (2019) reduces the entire field of statistics to just three questions: "How big is it? What difference does it make? Are you sure it's not just dumb luck?" While we must be careful to avoid overinterpreting the labels we assigned to principal components, it would seem no coincidence that the two major types of quantitative clauses answer the first two of these three questions.

News stories required a wide range of quantitative reasoning, with clear jumps rather than a smooth increase. Economy and health stories in our data set typically

required more quantitative reasoning than science and politics stories did. While our data set might not fully represent the news landscape, this result fits the first two authors' intuition as frequent news readers: we have both observed that economic and health reporting tend to rely heavily on official statistics, particularly in breaking-news contexts, while science reporting tends to treat findings as factual rather than probabilistic. Our journalist authors also note that science reporting is frequently done on slower timelines, giving reporters more time to interpret. All the same, the amount and types of quantification present in a story were not sufficient information to predict what category of news it belonged to, and we did not detect differences in quantitative content between legacy and online-first outlets or text and video stories (**RQ1**).

This is not surprising: broad categories like Economy, Health, Science, and Politics are not mutually exclusive. Furthermore, any such category includes a huge variety of topics, and for any specific topic, there are many ways a story can be presented. For example, a Politics story on how candidates are polling may discuss details of how the poll was conducted, may look at how each candidate's numbers have changed over the past few months, and may make projections on how likely a specific candidate is to win, each increasing the amount of quantitative information. However, it could also focus on potential reasons a particular candidate gained or lost support, or what strategies the candidate's campaign is planning to use to increase support, details which do not necessarily include quantification. We theorize that the level of quantification used to report on any given story topic may vary at the outlet level.

Previous efforts to evaluate the presence of quantification in news texts have typically focused on a single news outlet or a relatively small set of outlets. Maier (2002) looked at one newspaper, while Cushion et al. (2017) focused on the BBC across mediums, with additional information from commercial TV news, and these studies have not focused on the relationship between the small scale (e.g., a single statistic) and the larger scale (e.g., what is needed to understand the whole story). While their data sets were of the same size as our corpus or larger (>1,000 stories), this previous work had a narrower focus both in terms of sources and categories of quantification.

With an eye toward future studies, we note two limitations of the present study: the timing of data collection and the amount of data collected.

As noted in the Introduction, we collected this corpus in February 2020, as COVID-19 was receiving increasing coverage in U.S. media. The coverage had not yet started reporting daily case counts, job losses, and the other eventual consequences of the pandemic. Nonetheless, news sources had started reporting on projections of those quantities, which may have skewed our results. That said, the topics with the highest quantitative reasoning demands in our corpus, economy and

health reporting, tend to rely heavily on statistics even without an impending pandemic (for economics see Jensen 1987; for health see Reyna et al. 2009).

Our corpus included 230 of the top news stories (as ranked by Google News and analytics from PBS NewsHour), produced by 74 distinct outlets and covering four topic areas. Previous research has relied on larger numbers of stories (e.g., Cushion et al. 2017), but with a narrower range of topics or producing outlets. The size of our well-balanced corpus allowed for comparisons between topic areas, producing outlet types, and mediums. It also allowed us to extract reliable dimensions of quantitative knowledge requirements and quantitative content types. From these dimensions, we further extracted five clusters of news stories with similar quantitative profiles. Three of these clusters met all our criteria for reliability, including the number and size of the clusters, as well as statistics used for model selection and cross validation. A future study with a larger number of stories may corroborate the two additional clusters of stories we found.

Increasing the size of the corpus may also help with our secondary goal for the present analysis: training a machine-coding algorithm. Such an algorithm could "read" mass quantities of news stories and classify them based on the five quantitative profiles that defined our news clusters. These classifications would allow researchers to compare stories from different times and different sources, among other comparisons. Most important, these classifications would allow researchers to draw inferences about the quantitative knowledge and reasoning of audiences from the quantitative profiles of stories that dominate their news habits. Thus far, we have found that the number of stories is too small to provide linguistic patterns that are reliably associated with the classification codes.

Whether stories are human-coded or machine coded, examining the relationship between the quantitative knowledge of audiences and their news habits is a research priority for our team of journalists and social scientists. Audiences rely on the news to make informed decisions about their health, their finances, and their political behavior, among other essential decisions. In a probabilistic world where decisions depend on quantitative reasoning, the news needs to promote and support effective quantitative reasoning (cf. Barchas-Lichtenstein et al. 2021).

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. DRL-1906802. The authors are particularly grateful to colleagues who helped us think through this work, many of whom read earlier drafts of this paper. They include project evaluators Jim Hammerman and Eric Hochberg at TERC; project advisors Jim Corter, Danny Bernard Martin, Caitlin Petre, Jonathan Stray, Nikki Usher, and Darryl Yong; Rupu Gupta, John Fraser, and Nicole LaMarca at Knology; and Travis Daub, Vanessa Dennis, Molly Finnegan, Erica

Hendry, Chloe Jones, Megan McGrew, Miles O'Brien, and James Williams at *PBS NewsHour*.

References

Ancker, Jessica. 2020. "The COVID-19 Pandemic and the Power of Numbers." *Numeracy* 13, no. 2. https://doi.org/10.5038/1936-4660.13.2.1358

Attaway, Bennett, John Voiklis, and Jena Barchas-Lichtenstein. 2020. "Numbers in the News: Margin of Error." September 20, 2020. https://knology.org/article/numbers-in-the-news-margin-of-error/.

Bandy, Jack, and Nicholas Diakopoulos. 2020. "Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News." In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14: 36–47.

Barchas-Lichtenstein, Jena, John Fraser, Patti Parson, Rebecca Joy Norlander, Julia Griffin, Nsikan Akpan, Travis Daub, et al. 2020. "Negotiating Genre and New Media for STEM News." *Journalism Practice* 14, no. 6: 643–63. https://doi.org/10.1080/17512786.2019.1631711

Barchas-Lichtenstein, Jena, John Voiklis, Laura Santhanam, Nsikan Akpan, Shivani Ishwar, Bennett Attaway, Patti Parson, and John Fraser. 2021. "Better News about Math: A Research Agenda." *Numeracy* 14, no. 1. https://doi.org/10.5038/1936-4660.14.1.1377

Benton, Adia. 2016. "Risky Business: Race, Nonequivalence and the Humanitarian Politics of Life." *Visual Anthropology* 29, no. 2: 187–203. https://doi.org/10.1080/08949468.2016.1131523

Benton, Adia, and Kim Yi Dionne. 2015. "International Political Economy and the 2014 West African Ebola Outbreak." *African Studies Review* 58, no. 1: 223–36. https://doi.org/10.1017/asr.2015.11

Best, Joel. 2020. "COVID-19 and Numeracy: How about Them Numbers?" *Numeracy* 13, no. 2. https://doi.org/10.5038/1936-4660.13.2.1361

Bhatti, Yosef, and Rasmus Tue Pedersen. 2016. "News Reporting of Opinion Polls: Journalism and Statistical Noise." *International Journal of Public Opinion Research* 28, no. 1: 129–141. https://doi.org/10.1093/ijpor/edv008

Briggs, Charles L., and Mark Nichter. 2009. "Biocommunicability and the Biopolitics of Pandemic Threats." *Medical Anthropology* 28, no. 3: 189–198.

Briggs, Charles L. 2011. "On Virtual Epidemics and the Mediatization of Public Health." *Language & Communication* 31, no. 3: 217–228. https://doi.org/10.1080/01459740903070410

Carey, Benedict, and James Glanz. 2020. "Hidden Outbreaks Spread through U.S. Cities Far Earlier than Americans Knew, Estimates Say." *The New York*

Times, April 23, 2020. https://www.nytimes.com/2020/04/23/us/coronavirus-early-outbreaks-cities.html.

Clark, James M., and Allan Paivio. 1991. "Dual Coding Theory and Education." *Educational Psychology Review* 3, no. 3: 149–210. https://doi.org/10.1007/BF01320076

Crespi, Irving. 1980. "Polls as Journalism." *Public Opinion Quarterly* (1980): 462–476. https://doi.org/10.1086/268617

Cushion, Stephen, Justin Lewis, and Robert Callaghan. 2017. "Data Journalism, Impartiality and Statistical Claims." *Journalism Practice* 11, no. 10: 1198–1215. https://doi.org/10.1080/17512786.2016.1256789

Dunwoody, Sharon, Friederike Hendriks, Luisa Massarani, and Hans Peter Peters. 2018. "How Journalists Deal with Scientific Uncertainty and What That Means for the Audience." Presented at *15th International Public Communication of Science and Technology Conference, Dunedin, New Zealand, 2018*. https://pcst.co/archive/conference/paper/download/225

Figdor, Carrie. 2017. "(When) Is Science Reporting Ethical? The Case for Recognizing Shared Epistemic Responsibility in Science Journalism." *Frontiers in Communication* 2. https://doi.org/10.3389/fcomm.2017.00003

Fleerackers, Alice, Michelle Riedlinger, Laura Moorhead, Rukhsana Ahmed, and Juan Pablo Alperin. 2021. "Communicating Scientific Uncertainty in an Age of COVID-19: An Investigation into the Use of Preprints by Digital Media Outlets." *Health Communication* (January 2021): 1–13. https://doi.org/10.1080/10410236.2020.1864892

Fraley, Chris, and Adrian E. Raftery. 2002. "Model-based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association* 97, no. 458: 611–631. https://doi.org/10.1198/016214502760047131

Gal, Iddo. 2002. "Adults' Statistical Literacy: Meanings, Components, Responsibilities." *International Statistical Review* 70, no. 1: 1–25. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x

Gal, Iddo, and Irena Ograjenšek. 2017. "Official Statistics and Statistics Education: Bridging the Gap." *Journal of Official Statistics* 33, no. 1: 79–100. https://doi.org/10.1515/jos-2017-0005

Gao, Jie, and James E. Corter. 2020. "Describing and Comprehending Change in Quantitative Information." Paper presented at *42nd Annual Meeting of the Cognitive Science Society, virtual, July 29–August 1, 2020*.

Haim, Mario, Andreas Graefe, and Hans-Bernd Brosius. 2018. "Burst of the Filter Bubble?" *Digital Journalism* 6, no. 3: 330–43. https://doi.org/10.1080/21670811.2017.1338145

Hart, Tabitha, and Peg Achterman. 2017. "Qualitative Analysis Software (ATLAS.ti / Ethnograph / MAXQDA / Nvivo)". In *The International*

Encyclopedia of Communication Research Methods, edited by Jörg Matthes, Christine S. Davis, and Robert F. Potter. John Wiley & Sons. https://doi.org/10.1002/9781118901731.iecrm0194

Hinnant, Amanda, and María E. Len-Ríos. 2009. "Tacit Understandings of Health Literacy: Interview and Survey Research with Health Journalists." *Science Communication* 31, no. 1: 84–115. https://doi.org/10.1177/1075547009335345

Hope, Wayne. 2011. "Global Financial Crisis: Time, Communication and Financial Collapse." *International Journal of Communication* 4: 649–669

Irwin, Neil. 2020. "The Economic Data Is about to Get Weird." *The New York Times*, April 18, 2020. https://www.nytimes.com/2020/04/18/upshot/economic-data-distorted-coronavirus.html (accessed April 21, 2020.)

Jensen, Klaus Bruhn. 1987. "News as Ideology: Economic Statistics and Political Ritual in Television Network News." *Journal of Communication* 37, no. 1: 8–27. https://doi.org/10.1111/j.1460-2466.1987.tb00964.x

Karaali, Gizem, Edwin Villafane-Hernandez, Jeremy Taylor, and Pomona College. 2016. "What's in a Name? A Critical Review of Definitions of Quantitative Literacy, Numeracy, and Quantitative Reasoning." *Numeracy* 9, no. 1. https://doi.org/10.5038/1936-4660.9.1.2

Kleinnijenhuis, Jan, Friederike Schultz, Dirk Oegema, and Wouter van Atteveldt. 2013. "Financial News and Market Panics in the Age of High-frequency Sentiment Trading Algorithms." *Journalism* 14, no. 2: 271–91. https://doi.org/10.1177/1464884912468375

Koetsenruijter, A. Willem M. 2011. "Using Numbers in News Increases Story Credibility." *Newspaper Research Journal* 32, no. 2: 74–82. https://doi.org/10.1177/073953291103200207

Kohl, Patrice Ann, Soo Yun Kim, Yilang Peng, Heather Akin, Eun Jeong Koh, Allison Howell, and Sharon Dunwoody. 2016. "The Influence of Weight-of-Evidence Strategies on Audience Perceptions of (Un)Certainty When Media Cover Contested Science." *Public Understanding of Science* 25, no. 8: 976–91. https://doi.org/10.1177/0963662515615087

Lee, Angela M., and Hsiang Iris Chyi. 2015. "The Rise of Online News Aggregators: Consumption and Competition." *International Journal on Media Management* 17, no. 1: 3–24. https://doi.org/10.1080/14241277.2014.997383

Lehmkuhl, Markus, and Hans Peter Peters. 2016. "Constructing (Un-)Certainty: An Exploration of Journalistic Decision-Making in the Reporting of Neuroscience." *Public Understanding of Science* 25, no. 8: 909–926. https://doi.org/10.1177/0963662516646047

Mahalanobis, Prasanta Chandra. 1936. "On the Generalized Distance in Statistics." *Proceedings of the National Institute of Sciences of India* 2, no. 1 (January): 49–55.

Maier, Scott R. 2002. "Numbers in the News: A Mathematics Audit of a Daily Newspaper." *Journalism Studies* 3, no. 4: 507–19. https://doi.org/10.1080/1461670022000019191

Manski, Charles F. 2015. "Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern." *Journal of Economic Literature* 53, no. 3 (2015): 631–653. https://doi.org/10.1257/jel.53.3.631

McConway, Kevin. 2016. "Statistics and the Media: A Statistician's View." *Journalism* 17, no. 1: 49–65. https://doi.org/10.1177/1464884915593243

Nechushtai, Efrat, and Seth C. Lewis. 2019. "What Kind of News Gatekeepers Do We Want Machines to Be? Filter Bubbles, Fragmentation, and the Normative Dimensions of Algorithmic Recommendations." *Computers in Human Behavior* 90 (January): 298–307. https://doi.org/10.1016/j.chb.2018.07.043

Nguyen, An, and Jairo Lugo-Ocando. 2016. "The State of Data and Statistics in Journalism and Journalism Education: Issues and Debates." *Journalism* 17, no. 1: 3–17. https://doi.org/10.1177/1464884915593234

O'Connor, Cliodha, and Helene Joffe. 2020. "Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines." *International Journal of Qualitative Methods*, 19: 1–13. https://doi.org/10.1177/1609406919899220

Paletz, David L., Jonathan Y. Short, Helen Baker, Barbara Cookman Campbell, Richard J. Cooper, and Rochelle M. Oeslander. 1980. "Polls in the Media: Content, Credibility, and Consequences." *Public Opinion Quarterly* 44, no. 4 (1980): 495–513. https://doi.org/10.1086/268619

Pearson, Joel, and Stephen M. Kosslyn. 2015. "The Heterogeneity of Mental Representation: Ending the Imagery Debate." *Proceedings of the National Academy of Sciences* 112, no. 33: 10089–10092. https://doi.org/10.1073/pnas.1504933112

Pew Research Center. 2015. "Digital: Top 50 Online News Entities (2015)." https://web.archive.org/web/20160310155551/http://www.journalism.org/media-indicators/digital-top-50-online-news-entities-2015/

Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life.* Princeton University Press. https://doi.org/10.1515/9781400821617

Portilla, Idoia. 2016. "The Inclusion of Methodological Information in Poll-based News: How Do Spanish Newspapers Conform to Professional Recommendations and Legal Requirements?" *Journalism* 17, no. 1: 35–48. https://doi.org/10.1177/1464884915593239

Reyna, Valerie F., Wendy L. Nelson, Paul K. Han, and Nathan F. Dieckmann. 2009. "How Numeracy Influences Risk Comprehension and Medical

Decision Making." *Psychological Bulletin* 135, no. 6: 943–73. https://doi.org/10.1037/a0017327

Schmandt-Besserat, Denise. 1992. *Before Writing, Vol. I: From Counting to Cuneiform.* University of Texas Press.

Soroka, Stuart N., Dominik A. Stecula, and Christopher Wlezien. 2015. "It's (Change in) the (Future) Economy, Stupid: Economic Indicators, the Media, and Public Opinion." *American Journal of Political Science* 59, no. 2: 457–474. https://doi.org/10.1111/ajps.12145

St. John, Mark F. 1992. "The Story Gestalt: A Model of Knowledge-intensive Processes in Text Comprehension." *Cognitive Science* 16, no. 2: 271–306. https://doi.org/10.1207/s15516709cog1602_5

Tufféry, Stéphane. 2011. *Data Mining and Statistics for Decision Making.* Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470979174

United States Census Bureau. 2019. "All of Statistical Analysis in 3 Minutes." https://www.census.gov/programs-surveys/sis/resources/videos/all-of-statistical-analysis.html

Utts, Jessica. 2003. "What Educated Citizens Should Know About Statistics and Probability." *The American Statistician* 57, no. 2: 74–79. https://doi.org/10.1198/0003130031630

Vrieze, Scott I. 2012. "Model Selection and Psychological Theory: A Discussion of the Differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)." *Psychological Methods* 17, no. 2: 228. https://doi.org/10.1037/a0027127

Weaver, David A., and Bruce Bimber. 2008. "Finding News Stories: A Comparison of Searches Using Lexisnexis and Google News." *Journalism & Mass Communication Quarterly* 85, no. 3: 515–530. https://doi.org/10.1177/107769900808500303

Yarnall, Louise, SRI International, Michael Andrew Ranney, and University of California, Berkeley. 2017. "Fostering Scientific and Numerate Practices in Journalism to Support Rapid Public Learning." *Numeracy* 10, no. 1. https://doi.org/10.5038/1936-4660.10.1.3

YouGov. 2021. "The Most Popular News Websites in America." https://today.yougov.com/ratings/media/popularity/news-websites/all (accessed March 2, 2021).

Zwick, William R., and Wayne F. Velicer. 1986. "Comparison of Five Rules for Determining the Number of Components to Retain." *Psychological Bulletin* 99, no. 3 (1986): 432. https://doi.org/10.1037/0033-2909.99.3.432

References from the Data Set

Associated Press. 2020. "WATCH: WHO Says Data Shows Decline in New COVID Cases." *Associated Press*, syndicated by *PBS NewsHour*, February 20, 2020. https://www.pbs.org/newshour/health/watch-who-says-data-shows-decline-in-new-covid-19-cases

Austin, Paige. 2020. "As Outbreak Peaks, More Worry about Flu than Coronavirus: Poll." *Patch Hollywood*, February 21, 2020. https://patch.com/california/hollywood/outbreak-peaks-more-worry-flu-coronavirus-poll

Bloomberg. 2020. "This Year's Flu Vaccine Stopped Many of the Worst Cases, CDC Says." *Bloomberg QuickTake News*, February 21, 2021. https://www.youtube.com/watch?v=XznsyQg5p_Y

Boak, Josh. 2020. "White House Report Says Economy Accelerated under Trump." *Associated Press*, syndicated by *AOL*, February 20, 2020. https://www.aol.com/article/finance/2020/02/20/white-house-report-says-economy-accelerated-under-trump/23931813/

Brown, Troy. 2020. "Influenza Activity in US Remains High, and Still Rising." *Medscape*, February 19, 2020. https://www.medscape.com/viewarticle/925443

Cassella, Carly. 2020. "We've Vastly Underestimated How Much Methane Humans Are Spewing into the Atmosphere." *ScienceAlert*, February 19, 2020. https://www.sciencealert.com/we-ve-vastly-miscalculated-how-much-methane-humans-are-spewing-into-the-atmosphere

Colen, Aaron. 2020. "Coronavirus Is Much Deadlier than the Common Flu Despite Comparisons, New Analysis Shows." *The Blaze*, February 18, 2020. https://www.theblaze.com/news/coronavirus-deadlier-than-common-flu

Cox, Jeff. 2020. "America's Manufacturing Recession Looks Like It Could Be Over." *CNBC*, February 20, 2020. https://www.cnbc.com/2020/02/20/americas-manufacturing-recession-looks-like-it-could-be-over.html

Enten, Harry. 2020. "Trump's 2020 Position Is Improving." *CNN*, February 20, 2020. https://www.cnn.com/2020/02/23/politics/donald-trump-2020-poll-of-the-week/index.html

Fox News. 2020. "Obama Claims Credit for 'Longest Streak of Job Creation' in US History." *FOX* News, February 17, 2020. https://www.youtube.com/watch?v=-bPsyrIKm5s

Landler, Mark. 2020. "Spreading across Continents, a Lethal Virus Tests a Fraying Social Order." *The New York Times*, February 24, 2020. Retrieved from: https://www.nytimes.com/2020/02/24/world/europe/coronavirus-global-response.html

Lavoie, Denise. 2020. "National Shortage of Forensic Nurses Frustrates Rape Victims." *Associated Press,* syndicated by *PBS NewsHour*. February 19, 2020. https://www.pbs.org/newshour/health/national-shortage-of-forensic-nurses-frustrates-rape-victims

Marketwatch. 2020. "Asian Markets Mixed after China Cuts Loan Prime Rate." *Marketwatch*, February 19, 2020. https://www.marketwatch.com/story/asian-markets-mixed-after-china-cuts-loan-prime-rate-2020-02-19

MacAneny, DJ. 2020. "Total Delaware Flu-related Deaths for the 2019–2020 Season Rise to 9." *WDEL*, February 21, 2020. https://www.wdel.com/news/total-delaware-flu-related-deaths-for-the---season/article_c2ef238e-5500-11ea-86de-938c39dc1ac5.html

McCormick, Emily. 2020. "Coronavirus, Consumer Sentiment, GDP, Retail Earnings: What to Know in the Week Ahead." *Yahoo! Finance*, February 23, 2020. https://finance.yahoo.com/news/coronavirus-consumer-sentiment-gdp-retail-earnings-what-to-know-in-the-week-ahead-212252607.html

Miller, Rich. 2020. "White House Admits that Trump Trade Stance Did Depress Economy." *Bloomberg*, February 20, 2020. https://www.bloomberg.com/news/articles/2020-02-20/white-house-admits-that-trump-trade-stance-did-depress-economy

Newburger, Emma. 2020. "IMF Lowers Global Growth Forecast, Cases Surge in South Korea." *CNBC*, February 22, 2020. https://www.cnbc.com/2020/02/22/coronavirus-live-updates-imf-lowers-global-growth-forecast.html

Olson, Jeremy. 2020. "Flu Shows Second Act in Minnesota in 'a Particularly Bad Year for Kids.'" *Minneapolis Star-Tribune*, February 20, 2020. http://www.startribune.com/flu-shows-second-act-in-minnesota/568050432/

Reuters. 2020. "Sexuality and Gender Identity May Be Risk Factors for Skin Cancer." *Reuters*, syndicated by *NBC News*, February 20, 2020. https://www.nbcnews.com/feature/nbc-out/sexuality-gender-identity-may-be-risk-factors-skin-cancer-n1138841

Robb, Greg. 2020. "Philly Fed Manufacturing Index Jumps to Highest Level in Three Years in February." *Marketwatch*, February 20, 2020. https://www.marketwatch.com/story/philly-fed-manufacturing-index-jumps-to-highest-level-in-three-years-in-february-2020-02-20

Tappe, Anneken. 2020. "American Paychecks Just Aren't Getting Much Bigger—Unless You're Rich." *CNN Business*, February 20, 2020. https://www.cnn.com/2020/02/20/economy/wage-inequality-rising-2019/index.html

Appendix A: Clause-Level Codes (Full Version)

Code	Description	Keywords
Official statistics and official statistics organizations	Reference to official statistics ("data and diverse information products [made] available to keep policy-makers, various user groups, and the general public apprised of the current economic and social situation" (cf. Gal & Ograjenšek, 2017, 86), as well as any organization or agency whose mandate includes the publication of official statistics. This includes governmental and non-governmental polling organizations (e.g., those found at https://en.wikipedia.org/wiki/List_of_polling_organizations#United_States). This also includes vague or non-specified sources of data. Organizations that are purely statistical will always receive this code; other government associations only receive this code in association with a statistic.	"Government data," "government studies," "government report," "official data," etc., are all signals that an official statistics organization is involved. This code includes government organizations such as the Census Bureau, the CDC, and NOAA, as well as quasi-governmental non-profit organizations like the American Cancer Society, Red Cross, etc.
Comparison	Comparison of statistical quantities (including proportions, means, etc.) across populations or topics—refers to comparisons of one value to a different value. Includes comparisons to some norm or expected value, even where the base rate is left unspecified. Must refer to values that are quantifiable and quantified, even if the specific quantities are not made explicit. Correlations—that is, one quantity varying in a fixed relationship to another—also fall into this category. Change over time in a single quantitative value also falls into this category, as well as the lack of change over time, whether or not the base rate is specified. Must refer to a value that is quantifiable and quantified, even if the specific quantity is not made explicit. This code is distinct from Variability in that Variability covers all part-to-whole relationships, values being compared to a larger group that they are a part of. Comparison only includes relationships between distinct groups, not overlapping ones.	This code includes references to "high values" or "low values" compared to a specified or unspecified base rate. Also includes references to "increase," "decrease," and "trending" in a certain direction compared to a specified or unspecified base rate. A value being "like" or "unlike" another is also included in this category.
Proportion or Percentage	References to proportions that may or may not include explicit percentages. Also includes unquantified references to rates when these rates are clearly designated within the text as quantifiable.	Unquantified rates often are signaled by terms like "homelessness has grown" or "unemployment is down." References to percentage points fall into this category even when they are described simply as "points."
Central Tendencies and Exceptions	Includes references to averages—either means or medians—whether the type of average is explicit or implicit. Includes references to modes. Must refer to values that are quantifiable and quantified, even if the specific quantities are not made explicit. This category also includes outliers and exceptions from the norm, assuming that they imply a typicality or average even if not directly stated.	This category includes wording such as "the typical X," "the usual X," "a hallmark of X," "except for X," "the only X," etc. Most uses of the word "average" fall under this category. Do include "average" where it means the mode in conjunction with some sort of categorical variable rather than a numeric one, e.g., "the average American lives in a city."

Variability, Concentration, and Variation	Reference to a concentration or uneven distribution of a statistical phenomenon. This category differs from sampling & representativeness because variability & concentration are challenges for sampling and representativeness but are not themselves used to generalize to a larger population. This code differs from Comparison in that it includes part-to-whole relationships.	References to subgroups of a population (e.g., "white Americans") and their variation from the total population are typically included in this category. References to margin of error are considered part of this category. This code includes phrases like "the kind of place where . . ." Also includes phrases like "across-the-board," which imply a lack of variability.
Risk and Probability	References to risk, likelihood, or probability (e.g., of exposure to danger, harm, or loss; or future events or outcomes). Includes predictions and forecasts in scientific or political senses. This category must be about quantifiable statistical values, not the concept of probability in a more general sense. This category differs from Sampling because sampling refers to generalizing present data from measured values, while risk refers to projecting future data based on measured values.	The words "could" and "should" are often cues that a sentence contains probabilistic concepts. "Expecting" or "predicting" a result are also often part of a probabilistic statement.
Magnitude and Scale	References to scale, amount, or number of values being examined or assessed (cf. Yarnall and Ranney, 2017); includes raw numbers and sometimes approximate numbers. Small numbers are also "magnitude" if they're specific. Specific percentages are NOT magnitude.	Phrases like "8,000,000 people," "75 cases," "0.02 inches," "millions of Americans," or "thousands of years" signify magnitude.
Sampling, Representativeness, and Generalizability	Reference to research methods that use a sample population to generalize across a larger population, whether explicit or implicit. Note that "representativeness" is distinct from diversity or "representation," though the two can overlap in certain contexts. This code contrasts with Enumeration and refers specifically to the method by which the numbers were generated.	This code often applies where specific research, studies, or surveys are mentioned. References to "estimation" or "estimates" based on data are frequently considered part of this category.
Enumeration (and Inclusion/Exclusion Criteria)	Refers to research methods that use full enumeration, i.e., counting, rather than estimates derived through statistical inference. Only use this code if you are certain that this is the method by which the number was derived. Specific criteria for what IS or IS NOT counted goes under this code.	This category includes references to votes and the census.

Appendix B: Sample Coded Text

The following text is taken from the official transcript of a story that appeared on the PBS NewsHour on February 24, 2020, which is available in full at https://www.pbs.org/newshour/show/with-new-outbreaks-of-covid-19-are-we-on-the-precipice-of-a-pandemic

Text	Code(s)
Judy Woodruff: The virus that quarantined whole cities in China has now spread to new countries, and fears are growing.	Magnitude & Scale
Wall Street cratered today, as major indexes plunged more than 3 percent.	Comparison Proportion/Percentage
The Dow Jones industrial average lost over 1,000 points to close at 27,960.	Central Tendencies & Exceptions Comparison Magnitude & Scale
The Nasdaq fell 355 points.	Comparison Magnitude & Scale
And the S&P 500 dropped 111.	Comparison Magnitude & Scale
All of this amid encouraging signs inside China.	-
Amna Nawaz begins our coverage.	-
Amna Nawaz: Some factories across Shanghai were back in business Monday, as cases outside the epicenter of China's coronavirus outbreak fell to the lowest number in a month.	Comparison Variability, Concentration, & Variation
World Health Organization officials say the number of infected people in China has now peaked and leveled off.	Comparison Official Statistics & Official Statistics Organizations
But beyond China's borders, the virus, and concerns over its spread have picked up momentum.	-
There are now confirmed cases in at least 32 countries.	Magnitude & Scale Official Statistics & Official Statistics Organizations Research Methods