What do news readers want?*

Gregory J. Martin† Cameron Pfiffer‡ Shoshana Vasserman§

May 2026

Abstract

Using a novel dataset covering the complete history of individual-level web traffic and digital subscriptions from a major metropolitan newspaper in the United States between 2020 and 2024, we investigate consumers' willingness to pay for different categories of news content, with particular focus on the kinds of coverage believed to generate civic externalities. Our identification strategy relies on the quasi-random arrival of paywall events which force consumers to subscribe if they wish to continue reading. Using this variation, we estimate a model of consumer demand and construct the optimal staff allocation for the paper under different counterfactual revenue models: a fully subscription-based model and a fully ad-supported model. Our results suggest that readers are willing to pay for local reporting, and that measures of demand based only on time-use substantially underestimate the value of "hard" news coverage on topics like local politics and public health. However, digital subscription revenues alone are insufficient to cover staff costs even at the highest revenue-generating sections of the paper. We use our model to estimate the subsidy required to expand the newspaper's production of investigative coverage.

*We are grateful to an anonymous local news organization for making this project possible, as well as to Jai Singh for his phenomenal research assistance. We also thank seminar audiences at the CEPR/Bocconi Media Workshop, Princeton, the University of Arizona, NBER IO, the Utah Winter Business Economics Conference, and APSA for helpful comments and feedback.

†Stanford GSB. Email: gjmartin@stanford.edu.

‡Letta. Email: cpfiffer@gmail.com.

§Stanford GSB and the NBER. Email: svass@stanford.edu.

> We expect the newspaper to serve us with truth no matter how unprofitable the truth may be . . . Nobody thinks for a moment that he ought to pay for his newspaper. He expects the fountains of truth to bubble, but he enters into no contract, legal or moral, involving any risk, cost, or trouble to himself.
>
> Walter Lippmann (1922), *Public Opinion*

The feature of modern democratic politics that one of its essential inputs—information about the state of the world and about public policies—is provided by a profit-motivated press, to a public not necessarily keen on paying for it, has been recognized at least since Lippmann (1922). For much of the 20th century, however, the problems posed by this situation were obscured by the tremendous growth of the advertising business: local monopolies in advertising markets gave newspapers the freedom, and retained earnings, to invest in some reporting driven by public service rather than strictly commercial imperatives. The independent basis of support provided by advertising revenues allowed newspapers to break away from their former role as mouthpieces for political parties and establish norms of professionalism and objectivity in reporting (Gentzkow et al., 2007; Petrova, 2011).

But the dawn of the internet era brought the problems identified by Lippmann sharply back into focus. Competition from new digital competitors wiped out newspapers' advertising profits, driving sharp cuts in staff and coverage felt most acutely in politics-related news (Djourelova et al., 2025; Angelucci and Cagé, 2019). Digital distribution eroded the excludability of original reporting, further weakening the commercial incentives underlying the production of news (Cagé et al., 2020). Today, a steady drumbeat of layoffs and closures, at both legacy media and newer digital-only outlets, calls into question the long-term viability of the news-gathering function so essential to democracy.

A possible alternative to advertising as a basis of support for news organizations is a subscription-supported model, where readers pay directly for access to coverage. A precondition for the viability of such a model is that consumers are willing to pay for news—in money, and not just in the tax on attention taken by advertising. If commercial news outlets are to not just survive as businesses but also continue to serve their public function as information producers, their consumers' willingness to pay must be sufficient to cover the costs of uncovering and reporting the kinds of information essential for civic and political life.

There are reasons for skepticism that news readers' willingness to pay could be sufficiently high to satisfy this condition. Among scholars of mass communication, an influential view holds that obtaining political information is not a primary driver of news consumption for most consumers. A large part of the consumption of such content, in this view, derives not from intrinsic demand from readers but from supply conventions which bundle it together with other, more popular content. The dominant broadcast media technologies of the 20th century induced *incidental* exposure to political content by "entertainment-seeking" citizens not primarily interested in politics (Prior, 2007; Arceneaux and

Johnson, 2013; Durante et al., 2019). And the influence of journalists—whose professional reputations are built on the kind of long-form reporting that is taught in journalism schools and rewarded in journalism prize competitions—within news organizations led to the production of more investigative or accountability journalism than readers truly wanted (Hamilton, 2004).

Furthermore, the expansion of media choice enabled first by cable TV and later by the internet unbundled news content, allowing readers uninterested in political news or investigative journalism to opt out. And new measurement technologies gave management at media organizations direct visibility into the correlates of reader attention as captured by clicks and other browser metrics. Given these trends, many observers have raised concerns that news production might shift permanently in the direction of shorter "clickbait" articles that are produced at a fast pace to maximize traffic (Christin, 2020; Petre, 2021).

In this paper, we aim to directly measure news readers' tastes for politically relevant content. Is it true that (most) readers are primarily entertainment seekers who would avoid hard news if given the choice? Can accountability journalism, with its higher production costs, survive as the internet continues to unbundle the components of newspapers' traditional product? Will news outlets' shift to a "click-based" mode of evaluation (Christin, 2020) necessarily lead to the elimination of investigative reporting?

Our approach to these questions relies on granular data on digital traffic and subscription purchases from the online edition of a daily newspaper in a large US city. Despite the vast changes in the news market over the last thirty years, legacy metropolitan dailies like this one still produce a disproportionate share of coverage satisfying public information needs at the local level (Mahone et al., 2019; Hamilton, 2016).[1] We are able to track readers' decisions to read and to subscribe (or not) at the individual-by-article level, and can compare readers' willingness to pay for such socially valuable content compared to articles (sports coverage, for example) with mostly private benefits.

Unlike existing work (Fan, 2013; George and Waldfogel, 2006) which typically measures product attributes at the newspaper level using the share of newspaper staff allocated to different sections, we have access to rich article-level characteristics based on each article's full text, and can measure reader interest at the article level. We take advantage of the newspaper's policy of "paywalling" users—preventing them from reading more than a small number of articles in any rolling 30-day period without enrolling in a paid subscription plan—to identify readers' willingness to pay for different kinds of content. Specifically, we show that under conditional exogeneity of paywall events, the differential subscription rate between paywalled and non-paywalled users identifies reader willingness to pay for article attributes. Once we condition on users' baseline propensity to be paywalled (which is a

[1] Mahone et al. (2019) use the Federal Communications Commission's definition of "Critical Information Needs" (CIN) coverage. This category includes "information on local, regional, and county candidates at all levels of governance (p. 26)," among other types of information essential for civic and political life. Hamilton (2016) focuses on investigative reporting, using the definition proposed by the Investigative Reporters and Editors association (IRE): "Reporting... of matters of importance to readers, viewers or listeners. In many cases, the subjects of the reporting wish the matters under scrutiny to remain undisclosed (p. 13)."

function of users' frequency of visits to the site), residual variation in paywall events derives from small differences in the timing of visits or the specific article selected at the time of crossing the paywall threshold.

Building on this idea, we estimate a model of supply and demand for online news. Our data encompasses four years of articles and over 600 million article visits, with changing paywall criteria and promotional menus throughout. Our estimates indicate that the types of articles that drive visits are often quite different from those that motivate subscriptions. While visits span the portfolio of content available, subscription events are concentrated on opportunities to unlock access to articles related to areas of critical news coverage like public health, economics and local politics. After grouping the newspaper's staff writers into "beats" according to their typical output, we estimate reader willingness to pay for the average article written by writers on the Local News beat—who cover topics related to local politics, schools, and the local economy—to be roughly double that for the average article written by Entertainment staffers.

To connect these demand estimates to production incentives, we build and estimate a model of article production based on staffing assignments. In our model, the newspaper allocates staff across beats to maximize a revenue objective given a fixed staff budget. We allow for heterogeneity in the rate of production and in the distribution of quality in the articles they produce. Adding more journalists to any beat increases the number of articles that are written in that beat on any given day, with diminishing returns in the quality of the marginal article that is written.

Our model of supply and demand can inform two distinct areas of policy uncertainty. First, for news outlets, there is a live question of how best to restructure their business in the wake of the disruption experienced since the turn of the century. It is clear enough that the old approach is not viable, as print subscription and advertising revenues continue their steady decline. But newspapers have struggled to settle on a replacement. We examine the implications of three alternative strategies, which generate revenues respectively from digital advertising, from monthly all-access subscriptions, and from à la carte sales of access to individual articles. The model allows us to quantify how the optimal staff allocation would change under these different revenue objectives.

We find that a switch from traffic- to subscription-revenue-maximizing objectives would lead to reallocation of staff, broadly speaking, from soft news to hard news topics. The Local News and Business beats grow at the expense of Entertainment under the subscription objective, while the reverse is true under the traffic objective. À la carte has similar implications for staffing as subscription-based pricing, but even if the newspaper were to set the optimal à la carte price, this structure is unable to generate as much revenue as its current all-access subscription structure. The kinds of readers who would ever subscribe to a regional newspaper value access to the full bundle of newspaper content—value that is reflected in willingness to pay relatively high monthly subscription prices.

Second, for regulators and the public at large, the provision of politically relevant news coverage has the features of a public good. Media scrutiny generates external benefits in the accountability of

public officials (Snyder and Strömberg, 2010; Besley and Burgess, 2002) which accrue to all citizens, including non-readers. Our estimates of readers' demand and willingness to pay can inform regulators' understanding of whether media pluralism and the production of local reporting can be maintained under free-market conditions, and if not, the degree of subsidy that would be required to preserve these functions of the press.

Our results show that the sections of the newspaper that produce the coverage with the most positive social externalities, such as investigative reporting or coverage of public health issues, also generate the largest marginal subscription revenues at current staff levels. However, even the sections that generate the highest subscription revenues are net revenue negative once salary costs are accounted for. Increasing staff above current levels would push marginal net revenue from digital subscriptions further into the negative.

As a result, avoiding further staff reductions and declines in the production of socially valuable coverage is likely to require philanthropic or government funding. Our results are informative about the cost-effectiveness of such subsidies. Donors might, for instance, be interested in supporting the production of investigative journalism. On the margin, we estimate that the annual subsidy required to allow the newspaper to expand production of investigative articles without net revenue loss is around $11,000 per investigative piece per year.[2]

We proceed by first introducing our data and the context from which it was collected in Section 1. Section 2 discusses descriptive patterns in the data, both on the readership and the production side. Section 3 introduces our model of reader demand and our estimates of its parameters. Section 4 introduces the model of production and uses it to estimate newspaper responses to alternative revenue incentives. Section 5 concludes.

1 Background and Data

Our dataset comes from a metropolitan daily newspaper headquartered in a large US city.[3] The newspaper is typical of many legacy papers in the US in that it enjoyed a local monopoly during the print era, has since experienced a large decline in print subscription and display advertising revenue,[4] and has refocused on growing its digital subscription and advertising businesses. Like many US newspapers, it is also currently owned by a private-equity-controlled holding company.[5] The newspaper's audience is mostly but not exclusively regional: according to the Alliance for Audited Media, a slight

[2] We classify articles as investigative or not using the method of Turkel et al. (2021).

[3] Our data sharing agreement with the paper prevents us from disclosing the identity of the paper, or its raw subscription and revenue figures.

[4] Pew Research Center (2021) reports that aggregate weekday circulation of US daily newspapers declined from 44.4 million in 2011 to 24.3 million in 2020; estimated total newspaper advertising revenue dropped by more than $26B over this same period.

[5] See Ewens et al. (2023) for evidence on the prevalence and consequences of private-equity ownership of local newspapers.

majority of its online readers reside in the media market (Designated Market Area or DMA) in which the paper is headquartered.[6]

Digital subscriptions. Our data on prices and sales cover the newspaper's digital subscriptions, which currently account for around 40% of the newspaper's total subscribers.[7] Digital subscriptions offer subscribers unlimited access to articles published on the newspaper's website, in exchange for a monthly payment. Our data cover all digital subscription events in the period from January 2020 to December 2023. We observe the price paid and the terms of subscription, and can track the same reader before and after the subscription event.

Visits and article traffic. We observe individual visits to specific articles, which we can associate with unique and persistent reader identifiers. We group visits into user-sessions, corresponding to a contiguous sequence of visits by the same user. These data come from the newspaper's Google Analytics database, and cover all traffic to the site between January 2020 and December 2023. Traffic data includes the exact time of the visit, the referring website, and information about the length of time the user spent on the page and their engagement with the article (e.g. how far they scrolled down the page). For some analyses we aggregate traffic to the article level, counting the total visits or unique visitors occurring within 14 days of article publication.[8] We observe a total of about 1.2B user-sessions in our data and 605M article visits.[9]

Paywalls. The newspaper, like many others, uses a metering system to restrict access to its content. Non-subscribed users are limited to a small number, A_P, of free articles that they can read in any rolling T_P-day period.[10] The newspaper's system uses standard fingerprinting methods to construct stable identifiers of individual users and track their visits to the site. When a user's number of visits to an article page[11] in the past T_P days exceeds the limit, the user is presented with a pop-up that obscures the article content and presents subscription offers. If the user declines to subscribe, they are redirected to the home page and are unable to access the article. If they subscribe, they can continue

[6] DMAs are a definition of market boundaries originally defined by the Nielsen Company according to viewership of broadcast television stations and later adopted by the Federal Communications Commission in regulations governing ownership concentration and cross-ownership in media markets. They typically encompass one or more metropolitan counties and surrounding rural and exurban counties. There are 210 DMAs in the US.

[7] According to data from the Alliance for Audited Media's Media Intelligence Center reported in September 2022. This digital subscription share closely matches the 39% share of industry-wide advertising revenue coming from digital advertising in 2020 reported by Pew Research Center (2021). Both shares have been rising over time and are expected to continue rising.

[8] The time window ensures that later-published articles do not have mechanically lower traffic than earlier published articles.

[9] A large number of user-sessions visit only the home page or section header pages, and do not visit any article.

[10] A_P ranges between 2 and 4, while T_P is either 30 or 60. The newspaper changed one or both of these parameters at several points during our sample period.

[11] The newspaper's home page and navigation pages (such as section header pages) are unmetered and do not count towards the paywall cap.

to read the article, in addition to any other articles that they want to read in (at least) the next month.

Importantly for our purposes, the remaining meter count is completely opaque to the user; it is never displayed anywhere on the site. The number of free articles allowed (A_P) and the length of the rolling window (T_P) are also not disclosed to users, and these parameters changed at several points during our period of observation. In addition, technical glitches in the system at several points during our observation window temporarily disabled metering for some or all visitors. All of these features make it difficult or impossible for users to track their meter status, making the arrival of a paywall difficult to predict from the point of view of a reader. However, because the paywall is triggered after a certain number of articles are read and remains active thereafter until sufficient time has elapsed for the user's visits in the look-back window to drop below the threshold, readers who visit the site more often will on average hit paywalls more often. Our key identification assumption will be that *conditional on the user's reading history in the past 30 days*, paywall arrival—that is, the probability that a user i encounters a paywall when attempting to read article a—is independent of attributes of i or a.

In our data we observe more than 55M individual paywall events. We observe the exact time and date, a user identifier, and the article that the user was attempting to access when the paywall displayed. We also observe information about the subscription offers presented to the user, as described below.

Offers and menus. When activated, the paywall system presents users with a choice of a small number of subscription plans with different terms, which we refer to as a *menu*. For every paywall event, we observe a unique identifier ("menu code") corresponding to the particular menu of subscription offers presented to the user. We directly observe subscription prices and term lengths only for the specific offer that a user actually purchased from the menu (and thus, only for users who choose to subscribe). However, we can reconstruct the set of offers corresponding to a given menu by aggregating all of the offer purchases with the same menu code that were made on the same day or week. This process allows us to infer the choice set presented to any user with the same menu code, whether or not they made a purchase.[12]

The typical offer structure consists of a (usually low) *intro price* which applies for an *intro term* lasting one or more months; at the end of the intro term the introductory subscription converts to a monthly subscription at a (usually significantly higher) *renewal price* which automatically renews indefinitely each month until the user decides to cancel. This pricing structure is very common in the newspaper industry; see Miller et al. (2023). The first two offer parameters vary substantially, both across days due to the paper's use of temporary promotional sales, and across users on the same day, often due to experiments ("A/B tests") run by the newspaper. The last (renewal price) essentially takes one of two values depending on whether the user opts for an ad-free or standard renewal subscription; the former costs about 35% more than the latter. Empirically, a substantial fraction of subscribers end

[12] With the exception of any offers presented but never chosen by any user.

up paying the renewal price: roughly half of subscribing users continue their subscription for at least one month following the end of the intro term, and more than 25% are still subscribed 6 months after the end of the intro term (see Figure B.3 in the Appendix for details).

Readers who do not get paywalled can also subscribe, by clicking a link on any page on the newspaper's domain. This typically leads to a standard menu similar to that generated by a paywall. Notably, the newspaper often advertises a particular introductory offer through banner ads on commonly viewed pages like the home page. To account for these options, we also construct menus of offers for visits without paywall events. If no menu code is available for a given visit, we consider the most informative query parameter available in the visit's metadata and assign a menu based on the offers that were purchased with the same query parameters in the same day or week. In some cases, we observe more distinct offers than are possible (since the newspaper presents at most two digital offers at a time to a user). In these cases, we generate multiple feasible menus and assign them to visits probabilistically in proportion to the frequency with which the individual offers were observed.

Article content. We have full text, byline, and publication date information available for a set of 162,249 articles published by the newspaper between 2019 and 2023. Of these, 126,231 were published during the period covered by our traffic and subscription data (Jan 2020 - Dec 2023). We use the text and author information to construct several article features:

- **CIN category indicators.** We predict the article's likelihood of meeting each of eight "Critical Information Needs" (CIN) defined by Friedland et al. (2012) in a review commissioned by the Federal Communications Commission (FCC). These categories are Emergencies and Public Safety, Public Health, Education, Transportation, Environment and Planning, Economic Development, Civic Life, and Political Life. The method for constructing these predicted indicators, described in detail in Appendix A.1, relies on supervised learning using section labels, article full-text embeddings generated using the open-source model of Nussbaum et al. (2024), and counts of mentions of local places and politicians.
- **Additional non-CIN indicators.** We apply the same method as above to construct predictions that the article contains content in six additional categories which are not identified by the FCC as critical needs but which comprise a substantial fraction of the newspaper's production. These are Business, Entertainment, Real Estate, Sports, Things to do, and Opinion Columns.
- **In-house and wire service indicators**. Using the author information, we construct an indicator for the article being produced by a permanent staff member of the newspaper. This category excludes articles written by freelancers, by wire services, or by syndicated columnists. We also construct an indicator for the article being sourced from a wire service, such as the Associated Press or Reuters. Just under half of articles published in our sample period are written by a staff author, and about 32% are sourced from wire services.

- **Local reference indicator.** We searched the article's full text for the occurrence of a set of place names, sourced from the US Census, that are located inside the newspaper's home DMA. We construct a binary variable equal to one if there is any reference to a local place in the article, and zero otherwise. About 48% of articles published in our sample period refer to at least one local place name.

- **Investigative indicator.** Using the method of Turkel et al. (2021), we construct an "investigative score," a measure that uses the influence of an article on future articles in the corpus (see Blei and Lafferty 2006) to predict the article's similarity to past nominees of prizes honoring investigative journalism. We use a cutoff score of 0.1 to convert this continuous metric to a binary indicator; about 1.3% of articles published in our sample period exceed this threshold. See Turkel et al. (2021) for details of the method and validation of the measure.

User attributes. Our user attributes derive from the Google Analytics data. While we do not have information on user demographics, we do have the complete history of users' interaction with the site, which given our identification argument is the most important attribute on which we need to condition. We collapse users' recent article reading history into a scalar using principal components analysis (PCA). We extract the first principal component from a matrix consisting of the user's average reading depth of articles previously read, the total word count read by the user, the total number of previous articles visited, the word count of articles read in each of the previous six weeks, and the number of paywalls and registration walls[13] encountered in each of the previous six weeks. All of these variables are positively correlated, and the first principal component loads positively on all of them; we therefore consider it as a summary statistic of users' intensity of interest in the newspaper's website in the past month. The distribution of first dimension PCA scores (PC_1) is shown in Figure A.3 in the Appendix.

We use PC_1 to assign user-sessions into one of three discrete bins: user-sessions with PC_1 less than -0.5 are bin 1, between -0.5 and 0 are bin 2, and greater than 0 are bin 3. The score (and resulting bin assignment) can vary within user over time. Just about half of user-sessions are assigned to bin 1, about 28% to bin 2, and the remaining 22% to bin 3.

We are also able to track subscription status at the user-session level. We treat subscribers as a separate group from the non-subscribers, leading to four mutually exclusive categories of user: PC_1 bins 1,2,3, and subscribers.

Finally, for each user and each article visit, we estimate the user's probability of encountering a paywall when trying to read that article. The paywall probabilities are estimated according to a method described in more detail in Appendix A.2. Figure A.4 shows the distribution of estimated paywall probabilities for users in each of the three non-subscriber bins. The mean paywall probability

[13] A registration wall requires the user to create an account on the site (linked to an email account) to continue reading, but does not require payment.

by user type is, as expected, strictly increasing as we move from bin 1 to bin 2 to bin 3.

Conditional on encountering a paywall, users' propensity to subscribe also increases monotonically in our user bins. Table 1 shows the overall subscription rates of our user bins.[14] Bin 3 users are more than 100 times as likely to subscribe as those in bin 1, conditional on encountering a paywall.

Table 1: Average subscription rates by user bin. We compute the fraction of paywall events that convert to a paid subscription, by user bin. Rates are normalized so that bin 1 is 1.

Bin	Relative Subscription Rate
1	1
2	69
3	105

2 Descriptives and Reduced Form Evidence

We first present several stylized facts about the production of news content and the patterns of reading and subscribing by online consumers in our data.

2.1 Production

The newspaper offers a mixture of content across a variety of topical categories. Figure 1a shows histograms of the number of articles published on a single day from a given section. Some sections, like Sports, have high mean but also high variance, as coverage intensity tracks events like playoffs or championships. Other sections, like Advice, have much less variation in per-day article publication, with regular columns arriving on a predictable schedule. Figure 1b shows the number of distinct reporters observed writing articles primarily in the section; staff allocation aligns closely with the overall distribution of articles.

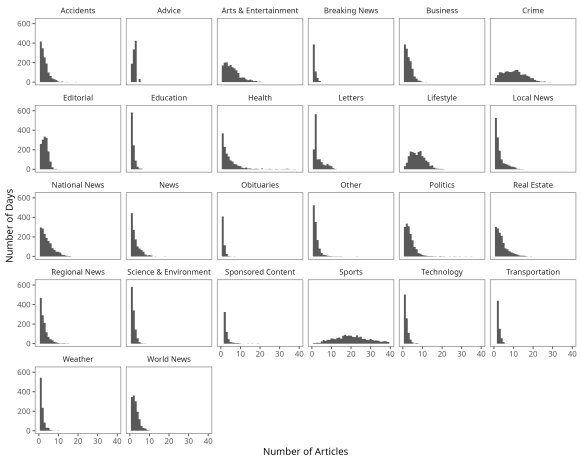
Our article feature classifications line up relatively well with the section breakdown. Figure 2 shows the average weight of each of our indicator features (each of which ranges from 0 to 1)[15] across all articles in the data, split by articles written by in-house staff (right panel) or non-staff (left panel). The most common features are Sports, Emergencies, Entertainment, Health, and Columns; as was evident in Figure 1b, the in-house staff are heavily weighted towards sportswriters and crime reporters. Figure A.2 in the Appendix shows that there is substantial variation in the mix of coverage that the paper offers over time, with Health coverage tracking the COVID-19 pandemic in 2020-2021 and Politics coverage tracking the electoral calendar. Figure A.1 in the Appendix shows the correlation matrix between each of our features.

[14] Relative to bin 1, to satisfy the condition of our data sharing agreement that prohibits revealing subscription numbers.

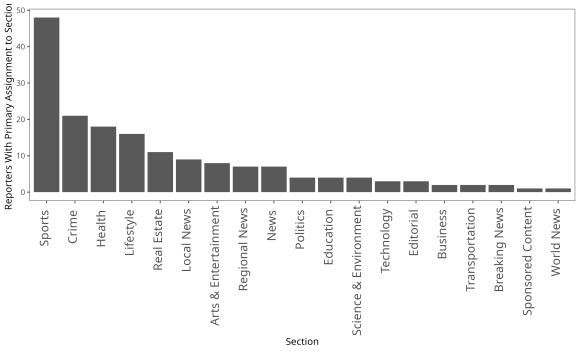
[15] The supervised-learning-based features are continuous, but their distributions are all highly bimodal with peaks near 0 and near 1.

Figure 1: Histograms of the number of articles published in a day per section, and bar graph of distinct reporters observed whose output is primarily in the section.

(a) Number of Articles Per Day



(b) Number of Dedicated Reporters



2.1.1 Defining journalistic "beats"

Our content data allows us to directly observe the newspaper's categorization of the articles it publishes into sections. However, there are over 100 distinct section labels in our data. The analysis in Figure 1 condenses these raw labels into a smaller set of 25 by collapsing small name variations, but nonetheless there remains substantial overlap between sections: authors who write in Local News may also be equipped to write in Regional News, for example. Indeed, we observe many authors who regularly write in multiple sections. We therefore use a clustering procedure to group authors who write similar kinds of articles together; we refer to the resulting clusters as "beats." The idea is to capture the broad categories of reporting (e.g., sports coverage versus political reporting) between which the newspaper's management could shift staff. We will use these beat definitions in both the demand estimation (Section 3) and in our supply model that we use to generate counterfactual production choices (Section 4).

To produce this clustering, we use the same set of article features introduced previously. For each staff author-month in the data,[16] we compute the average feature vector, which is just the average value across all published articles written by that author in that month.[17] The set of author-month averages is then input to a k-means clustering algorithm; we choose $k = 8$. We give the resulting beats a descriptive label using the typical sections in which authors in that beat publish. The beats can be described as Sports, Entertainment, Local News, Health, Business, Local Events, Editorial, and Crime. Clustering into 8 beats produces clearly distinct groupings: articles written by authors we assign to the Sports beat, for instance, have mean value of the Sports feature in articles they write of

[16] Articles written by wire services are not included in the clustering, as our purpose is to create groups of authors who work for the newspaper and could potentially be reassigned by management.

[17] For coauthored articles, we weight the article by the inverse of the number of authors in the computation of the average feature vector for each coauthor.

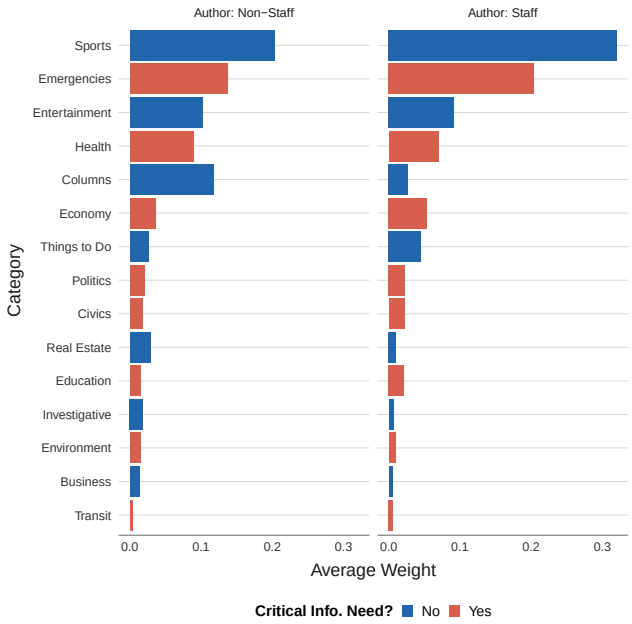


Figure 2: The average weight across all articles in the dataset, split by articles written by in-house staff (right) and by other authors such as wire services (left). Categories identified by the FCC as "Critical Information Needs" are highlighted in red.

0.95. See Figure A.5a for a visualization of each beat in terms of its typical feature composition.

We then compute the observed staffing level of each beat as the number of distinct authors that our clustering algorithm associates with the beat in each month. Figure A.5b displays the total staffing by beat by month. There are both month-to-month fluctuations in staffing and broad trends: the Health beat, for instance, spikes during the early months of the COVID-19 pandemic, and then declines. This change in staffing closely tracks the volume of article publication associated with the Health feature (see Figure A.2a).

2.2 Traffic and subscriptions

We next examine the visitor traffic and subscriptions that articles generate, according to our characterization of article features. We measure traffic as the number of unique visitors arriving at an article within 14 days of its publication. We measure subscription rates as the fraction of users who visited an article and were shown a paywall requiring them to subscribe in order to be able to read the article, who subscribed to some paid subscription plan in that session. Figure 3 shows the weighted mean values of visits and subscription rate, weighted by each feature.[18]

The figure makes clear that for almost all categories, articles written in-house generate substantially

[18] E.g., the visits value for the Columns category is computed by taking the weighted average of visits across all articles, where the weight is each article's value of the Columns feature: $\left(\sum_a Visitors_a w_a\right) / \left(\sum_a w_a\right)$ where w_a is the value of the Columns feature estimated for article a.

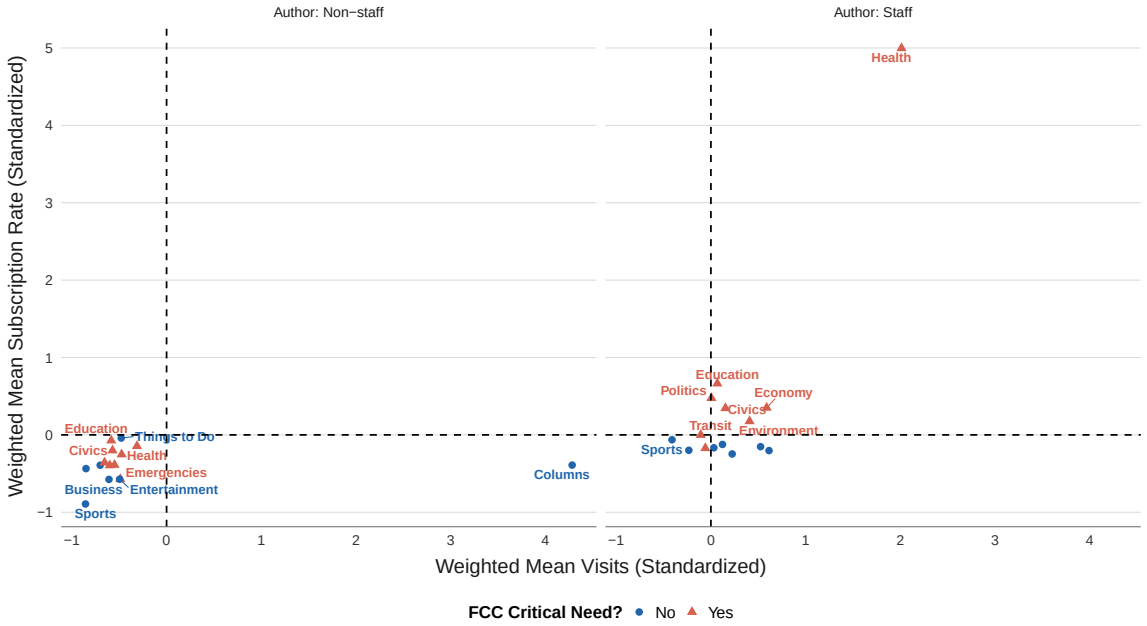


Figure 3: Visits versus subscriptions, by feature. The y-axis is the average value of the subscription rate for users paywalled on each article in the dataset, weighted by that article's value of the indicated feature. The x-axis is the average value of the number of unique visitors for each article in the dataset, weighted by that article's value of the indicated feature. Both axes are standardized by subtracting the mean and dividing by the standard deviation across all features.

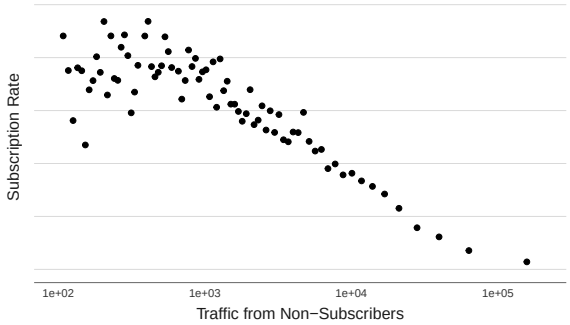
more traffic than outsourced articles. The important exception is the Columns category; by far the site's biggest traffic drivers are syndicated advice columns. Health and Economy-related articles also appear to generate substantially more traffic than the other categories.

On the subscription side, in-house articles also almost uniformly produce higher subscription rates than do outsourced articles. Within in-house articles, there is substantial variation in subscription rates. The pattern is generally that categories with higher social value (those identified by the FCC as Critical Information Needs) have higher conditional subscription rates than do articles covering less-critical topics like local sports. Health articles in particular are especially effective in generating subscriptions, perhaps because of the value of this type of coverage during the COVID-19 pandemic which occurred during our sample period. These univariate comparisons of means may miss the correlation structure of features across articles, but results presented in Appendix B.1 show that these conclusions hold in a regression setting as well.

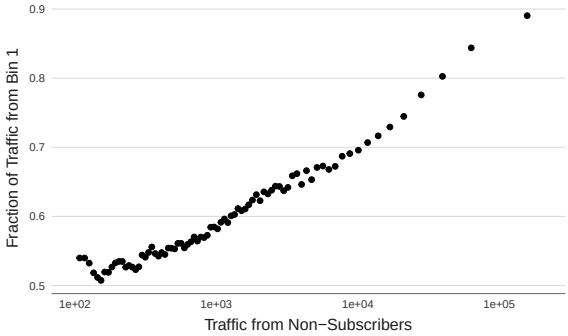
Our regression results are similar even if we condition not only on a coarse measure of total past reading (our discrete user bins) but also on a much finer measure that includes the type of articles that individual users previously read: see Figure B.4 in Section B.1.4. Hence, we interpret the differences in conditional subscription rates across article types as reflecting, at least to some extent, differences in willingness to pay *within person* and not just selection of different kinds of readers.

Figure 4: Average subscription rates versus total non-subscriber visitors to an article.

(a) Subscription rate.



(b) Traffic from lowest PC_1 bin.



Traffic shocks and subscription rates. Article content affects both traffic and subscription propensity, conditional on having arrived at the site and facing a paywall on some article. Our estimates of content effects on propensity to subscribe (Figure B.2) could be biased if some articles bring in visitors who have different baseline subscription propensities.[19]

Figure 4a shows that there is indeed a correlation between total traffic to an article and the fraction of paywalled readers who subscribe. Each point in the figure is the average of all articles within the same percentile of the logged traffic distribution. The figure shows that the correlation is fairly strong and negative, which implies that more popular articles draw in marginal readers who are generally less interested in subscribing. Figure 4b shows that this difference is well captured by our discrete user bins. As the total traffic to an article rises, the fraction of that traffic coming from the lowest bin (the least likely subscribers) rises substantially. Both patterns suggest a trade-off for the newspaper between optimizing for clicks or traffic versus optimizing for subscriptions. More popular articles tend to disproportionately draw in marginal readers who are unlikely to subscribe.

Price Sensitivity. Our discussion of subscription rates thus far has ignored prices, but as we described earlier, there was substantial variation in the terms offered to users over our sample period. We now investigate how sensitive users are to the offered prices and subscription terms. For a descriptive analysis, we consider a simplification that collapses the two primary dimensions of offer variation (intro price and intro length) into one: the average price per month over the intro term. This summary statistic is increasing in the intro price and decreasing in the length of the intro term, and so it succinctly characterizes the favorability of the intro offer terms. We match each paywall event to the menu that the user saw upon being paywalled and regress the share of paywalled menu views that resulted in a subscription on the lowest average price offered among all options in each menu.

Table 2 shows that users are responsive to price changes, and subscribe less often when offer terms are less favorable. The highest-engagement users (bin 3) are the most price sensitive, though this is

[19] Our use of PC_1 bins as a conditioning variables handles observable compositional change across articles, but there remains a possibility of within-bin differences in unobservables in the samples of users paywalled on different days.

probably due to the fact that bin 3's much higher baseline subscription levels make the correlation of subscription with price stronger.

Table 2: Price sensitivity by user bin.

		Subscribed		
Group	Full sample	1	2	3
	(1)	(2)	(3)	(4)
Price per Month	$-2.71 \times 10^{-5***}$	$-3.57 \times 10^{-7**}$	$-1.58 \times 10^{-5***}$	$-4.37 \times 10^{-5***}$
	(2.41×10^{-6})	(1.74×10^{-7})	(2.77×10^{-6})	(4.65×10^{-6})
Visit Date FEs	Y	Y	Y	Y
Observations	13,485,953	4,829,333	5,770,928	2,885,692
R^2	0.0006	0.0002	0.001	0.003

Notes: An observation is a distinct combination of day, article viewed, user bin, and menu of subscription plans offered. Observations are weighted by the total number of user-sessions in that day-by-article-by-menu-by-bin cell. The dependent variable is the fraction of user-sessions in the cell which purchased any paid subscription plan. The price variable is the lowest average price per month over the intro term available among the options in the menu shown to the user at paywall time. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

2.3 Subscribers' Reading

A third indication of consumer value comes from readers' change in habits after they subscribe. Whereas subscribers can read as many articles as they like without incurring additional costs, non-subscribers increase the likelihood that they will hit a paywall and be blocked from reading, the more that they read. Subscribers also likely face lower search costs of getting to the site, compared to non-subscribers. Both differences suggest that subscribers should be willing to read lower-value articles than non-subscribers.

Table 3 presents evidence for the decline in marginal cost associated with subscription. The table shows estimates from a regression of a variety of measures of engagement with the site on both extensive (number of visits to the site, number of articles read) and intensive margins (average depth into an article that the user reads, number of words read per visit) within a 120-day window centered on a subscription event. The regression includes user fixed effects plus an indicator for the user being subscribed. All of these measures increase significantly, usually by a factor of approximately 2, within user after subscription.

Figure 5 shows that subscribing also changes the composition of articles that users read. The average level of the CIN category measures (with the exception of Emergencies) in users' consumption bundle goes down, often significantly. The point estimate for the investigative indicator is also negative, although the base level is so low that the change is not significant. Per Table 3, these declines occur as the user reads more articles in total, suggesting that reading these kinds of articles becomes diluted by other content as total reading increases.

The changes in the composition of subscribers' reading, consistent with Figure B.2, indicate that CIN

Table 3: Site engagement before and after subscription.

	Visits (1)	Total Articles Read (2)	Words per Visit (3)	Average Depth (4)
Subscribed	8.65***	14.2***	336.3***	0.247***
	(0.099)	(0.153)	(2.63)	(0.0010)
User FEs	Y	Y	Y	Y
Observations	[redacted]	[redacted]	[redacted]	[redacted]
R^2	0.72	0.67	0.65	0.72
Mean dependent variable	11.5	13.9	407.7	0.25

Notes: The sample is all users subscribing during our period of observation. An observation is the average over a 60 day period, either pre- or post-subscription, of the indicated variable for a particular user. I.e., there are two observations per user in the sample. The dependent variables capture (1) the number of sessions (visits to a sequence of pages on the site separated by at least a half hour); (2) the total number of articles read; (3) the sum of the word count of articles visited multiplied by the user's reading depth for that article, and (4) the average fraction of the page to which the user scrolled on article pages. Observation counts are omitted per our data sharing agreement which prohibits releasing subscriber numbers. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

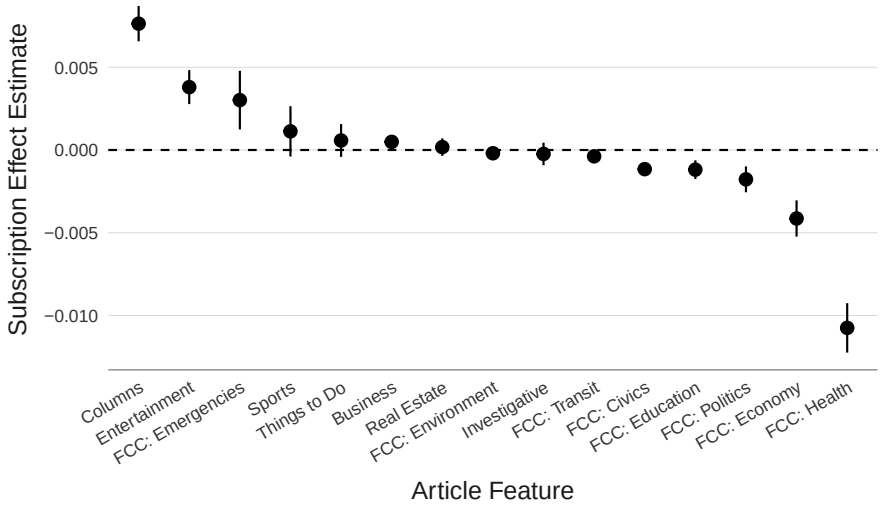
features are *positively* associated with users' valuation; when unsubscribed and reading opportunities are scarce, users allocate a greater fraction of their reading time to CIN content than they do when reading is unconstrained. Conversely, the average level of features like Columns and Entertainment goes up after subscription, indicating that these kinds of articles have lower private value.

3 Demand Model

In this section, we present a model of reader demand for online news. Our model combines the elements of demand highlighted in the previous section: (1) readers of different types may visit the newspaper's website more on days when more valuable content is available; (2) casual visitors to the newspaper read different types of content than subscribers; (3) readers are sensitive to prices on subscription offers; and (4) readers may be more likely to subscribe when they are blocked by a paywall from reading certain types of articles that they value.

Readers We consider a population of potential visitors to the newspaper website on each day. Readers vary in their exposure and baseline interest in the newspaper, as well as in their willingness to pay for a subscription. Readers' level of engagement with the newspaper may also vary over time, and change with their exposure. To account for this parsimoniously, we assume that reader preferences at each visit are characterized by the discrete type described in Section 1. We do not model dynamic decision-making by readers. If we observe the same reader over the course of several visits, we treat each visit as a distinct short-lived synthetic instantiation of the reader, characterized by the aggregate history of the real long-lived reader's interactions with the newspaper. That is, we denote each observation of a reader-visit pair as a new reader i with type x_i.

Figure 5: Content changes post-subscription. The figure shows point estimates and 95% confidence intervals of within-user changes in the average level of each feature in articles read by the user, after the user subscribes. Each coefficient is from a separate regression, where the outcome is the mean level of the indicated feature among all articles read by the user in the 60 day window pre- or post-subscription.



Articles Each day d, the newspaper publishes a new set of articles, varying in content based on current events, internal project timelines and journalistic capacity. Each article $a \in A_d$ is characterized by a vector of attributes θ_a, corresponding to the article features described in Section 1.

Reader Visits to Articles Readers are more likely to visit some types of articles than others, and are more likely to visit *any* article on the newspaper's website on certain days more than others. To capture this, we assume that readers who are not already subscribed to the newspaper arrive at an article a on day d according to a Poisson distribution parameterized as follows:

$$V_{a,d,x} \sim \text{Poisson}\left(\exp\left(\beta_x^V \cdot \theta_a + \xi_x^V + \frac{1}{||c(a)||}\sum_{c \in c(a)} \zeta_{c,w(d)}^V\right)\right). \tag{1}$$

Here, $V_{a,d,x}$ is the number of visits to article a on day d by a reader of type x. The Poisson rate for $V_{a,d,x}$ is given by a generalized regression function of article characteristics θ_a plus a reader-type intercept (ξ_x^V) and one or more beat-by-week fixed effects ($\zeta_{c,w(d)}^V$) under an exponential link. Beats are described in Section 1. The beat-week FE allows for reader interest in, say, articles written by the paper's sportswriters, to vary over time. An article can belong to multiple beats, denoted by the collection $c(a)$, if it is co-authored by authors we assign to different beats.

The vast majority (about 84%) of article visits occur within two days of an article's publication date. For parsimony, we therefore consider visits occurring three or more days post-publication to be "late arrivals." We group all such late visits for the same article together and model the total number of late arrivals to an article with a modified Poisson regression, replacing the reader-type intercept ξ_x^V

with a reader-type-specific late-arrival fixed effect ξ_x^{LV} and a control for the number of days $\tilde{D}_a$ that article a was available for reading during our entire panel window (of length $D = 1461$ days):

$$V_{a,x}^{LV} \sim \text{Poisson}\left(\exp\left(\xi_x^{LV} + \tau_x\left(1 - \frac{\tilde{D}_a}{D}\right) + \beta_x^V \cdot \theta_a\right)\right). \tag{2}$$

Finally, we model visits to pages without article characteristics (i.e., the newspaper home page) with a similar Poisson expression that replaces the article-specific terms with a generic "null article" intercept ξ_x^{NV}:

$$V_{\emptyset,d,x} \sim \text{Poisson}\left(\exp\left(\xi_x^{NV} + \xi_x^V\right)\right). \tag{3}$$

Paywalls Readers may be blocked from accessing an article's contents by a paywall for a number of reasons, including their use of an ad-blocker, using up their allocated free article count, or a server error. We assume that paywall events are random conditional on the user's observable characteristics (x_i) at the time of their visit, and are independent of the attributes of the article being visited.

Reader Values for Articles Readers only benefit from an article a if they *both* visit the article page *and* are able to read the article. That is, readers receive no value from visiting an article if they hit a paywall and are blocked from accessing the article's contents. We assume that a reader of type x who visits an article a and is able to read it receives an article-reading utility given by:

$$r_{a,d,x} = \beta_x^R \cdot \theta_a + \frac{1}{||c(a)||}\sum_{c \in c(a)} \zeta_{c,w(d)}^R.$$

Here, θ_a is the vector of article characteristics discussed before, while β^R and ζ^R are characteristic coefficients and beat $\times$ week fixed effects, respectively, as in the visit model. If the reader hits a paywall and is unable to read an article she visits, then she gets utility of 0 from that article unless she purchases a subscription offer.

Subscription Offers On each visit, readers receive a menu of promotional subscription offers. Each offer o consists of three components: an intro term duration T_o (in months) during which the reader would have unrestricted access to the newspaper's online content, an initial price p_o^0 which the reader pays for this period, a subscription tier (either ad-free or standard), and a renewal price p_o^1 which she will pay monthly to continue access if she does not cancel the subscription. To evaluate an offer o, readers consider both the immediate value of subscribing, and the net present value of being subscribed in the future. If the reader is blocked from reading an article a by a paywall, then the immediate value of subscribing is her value for the article $r_{a,d,x}$. If the reader is not paywalled, then there is no immediate value for subscribing and only the future ("bundle") value of being subscribed plays a role in her subscription decision.

We assume that readers are forward-looking but have imperfect ability to forecast the articles that will be published over the duration of their subscription offer: they project that articles published over the term of their subscription will, on average, match the average features of articles published by the newspaper in the week before and the week after their subscription date.[20] This assumption allows for the content that readers expect to be published while they are subscribed to change over time, either because of expected coverage of scheduled events such as elections and sporting events, or because they involve ongoing coverage of an evolving story.

The goal of the bundle value in our model is to both capture the effect of this predictable, time-varying news coverage, and to account for the difference in utility that subscription offers with different durations entail. In the data, we observe that the newspaper offers higher-priced offers with a longer subscription term, and that readers do purchase these subscriptions. To rationalize this behavior, we model the net present bundle value of an offer o considered on day d by a reader of type x as follows:

$$\begin{aligned} B(o,d,x) = & \mathbf{1}\{o \text{ is ad-free}\}\kappa_x^R + \\ & \sum_{t=1}^{T_o} \delta_x^{t-1} \left(b_x + \beta_x^R \cdot \bar{\theta}_{w(d)} + \zeta^R \cdot \bar{C}_{w(d)}\right) + \\ & \sum_{t=1}^{\max(15-T_o,0)} \delta_x^{T_o+t-1} \hat{\pi}_{x,t} \left(b_x + \beta_x^R \cdot \bar{\theta}_{w(d)} + \zeta^R \cdot \bar{C}_{w(d)} + \alpha_x p_o^1\right). \end{aligned} \tag{4}$$

The first component in (4) is the offer's subscription tier, which can be either "standard" or "ad-free". The newspaper advertises ad-free subscriptions as a higher-quality experience with faster load times and no distracting banner ads. We assume that this additional quality is additively separable from the article values through a parameter κ_x^R.

The second component in (4) is a sum over the intro term for offer o. Each month of subscription generates utility that is a linear function of the average feature vector $\bar{\theta}_{w(d)}$ and average beat $\times$ week indicator variables $\bar{C}_{w(d)}$ of all articles published in either the week containing the date on which the offer is evaluated ($w(d)$) or the week before or the week after. The three parameters in this function are b_x, a type-specific bundle value intercept, β_x^R, the same reader value coefficients used to compute the article-specific utility, and ζ^R, the same beat $\times$ week FE parameters used to compute the article-specific utility. Each month of utility is geometrically discounted at rate δ_x, a monthly discount factor specific to the user's type x.

The last component in (4) is a sum over the months following the end of the intro term, up to a maximum of 15. It contains two additional components. First, the subscription value in each month after the intro period ends is multiplied by the probability that the reader remains subscribed for t months following the end of the intro, $\hat{\pi}_{x,t}$, which we estimate directly from the data.[21] Second, the monthly subscription value also contains the monthly renewal price p_o^1, multiplied by the price

[20] See Appendix B.1.3 and Table B.1 for some reduced-form evidence supporting this choice.
[21] See Figure B.3 in the Appendix.

sensitivity parameter α_x.

Typical menus include both a standard and an ad-free option. However, various conditions may cause single offer menus to be shown. For instance, sessions in which an ad-blocker is detected are typically offered only the ad-free offer. In many cases, we can observe the specific menu that was shown to users along with relevant session information. In cases where the exact menu is unknown, we assume that the user saw one of the menus that was seen by other users with similar session details on the same week with a likelihood proportional to that menu's observed empirical frequency.

Subscription Propensity Readers faced with a menu of subscription offers can choose to purchase one of the offers, or not to subscribe. We assume that readers choose the option that maximizes their utility, given the immediate value of reading a paywalled article (if relevant), the net present value of being subscribed (given the duration of each offer), the offer price, a reader-type intercept ξ_x^R, and an idiosyncratic Type-I extreme value shock ε. In addition, we assume that paywall events trigger an additive reader-type specific shift in the utility from subscribing, denoted by η_x. Finally, we assume that reader utility is quasi-linear in prices, and subject to the price sensitivity parameter α_x.[22]

Putting these pieces together, let i denote a subscription opportunity for a reader of type x on day d visiting article a. The utility that the reader expects from purchasing an available offer o is given by:

$$u_{i,x,d,a}(o) = \alpha_x p_o^0 + \mathbf{1}\{i \text{ paywalled on } a\} \cdot \left(\eta_x^R + r_{a,d,x}\right) + B(o,d,x) + \xi_x^R + \varepsilon_{i,x,d,a,o}. \tag{5}$$

Let O_i denote the menu of offers given to the reader at opportunity i. Then the probability that the opportunity leads to a purchase of an offer $o \in O_i$ is given by:

$$\pi_{i,x,d,a}(o) = \frac{\exp(\alpha_x p_o^0 + \mathbf{1}\{i \text{ paywalled on } a\} \cdot \left(\eta_x^R + r_{a,d,x}\right) + B(o,d,x) + \xi_x^R)}{1 + \sum_{o' \in O_i} \exp(\alpha_x p_{o'}^0 + \mathbf{1}\{i \text{ paywalled on } a\} \cdot (\eta_x^R + r_{a,d,x}) + B(o',d,x) + \xi_x^R)}. \tag{6}$$

3.1 Estimation

Our demand model produces three sets of parameters to estimate, listed in Table 4. To estimate these parameters, we perform a two-step maximum likelihood procedure. In the first step, we estimate the parameters governing visits to different articles by maximizing the likelihood of the observed counts of article visits for each article on its publication day, one day after, two days after, and more than two days after publication given article features, as parameterized in equations (1) and (2). In the second step, we estimate the parameters governing the probability of subscription decisions. To do this, we maximize the likelihood of observed subscription decisions for every visit in our data, conditional on features of the visited article, the user type, the realization (or not) of a paywall event, the article

[22] Note that we impose that this is the same as the price sensitivity that readers apply to renewal prices in the bundle value calculation.

feature and beat-week fixed effects in the week before and after the visit, and the characteristics of offers presented in the menu shown to that user. This probability is given by equation (6).

Table 4: Demand Model Parameters

Group	Parameters
Visit parameters	Beat × week fixed effects: $\zeta^V_{c,w(d)}$ Article characteristic coefficients: β^V_x Reader-type intercepts: ξ^V_x Late arrival parameters: ξ^{LV}_x, τ_x Null article parameters: ξ^{NV}_x
Subscription parameters	Beat × week fixed effects: $\zeta^R_{c,w(d)}$ Article characteristic coefficients: β^R_x Reader-type intercepts: ξ^R_x Ad-free tier values: κ^R_x
Bundle and paywall parameters	Price coefficients: α_x Bundle value intercepts: b_x Monthly discount factors: δ_x Paywall intercepts: η^R_x

We calculate confidence intervals for all parameter estimates using the Bayesian bootstrap. For each bootstrap iteration, we re-weight the observations in our data using random draws from a Dirichlet distribution, and re-estimate the full two-step model. We repeat this procedure for one hundred iterations and use the 2.5th and 97.5th quantiles to compute the upper and lower 95% confidence bounds for the parameters.[23]

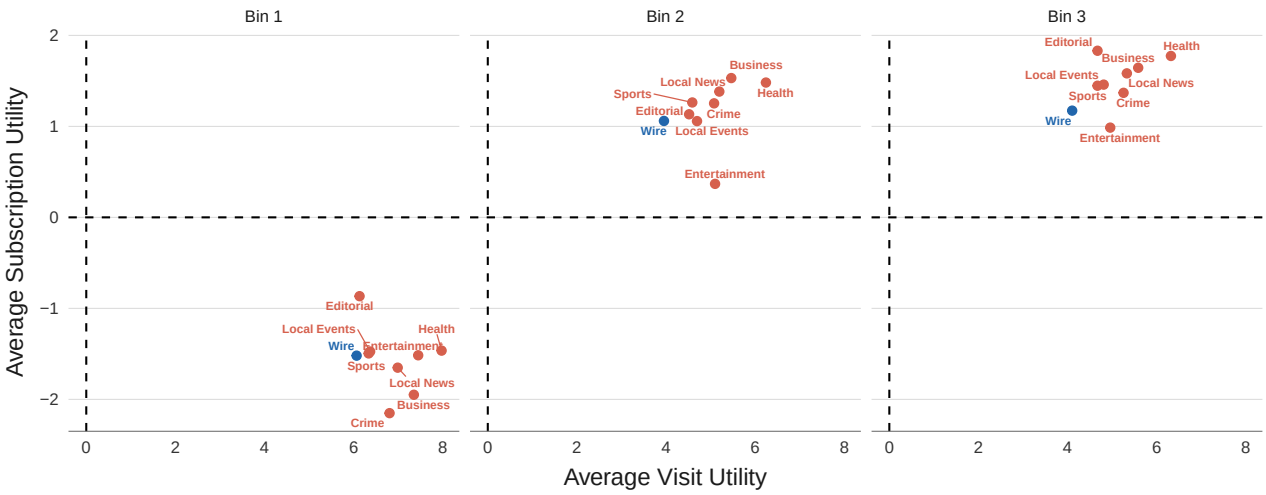
3.2 Results

We summarize our estimates of article-specific values in Figure 6: for each of the in-house author beats, plus an aggregate of articles sourced from wire services, we plot the average subscription utility (vertical axis) against the average visit utility (horizontal axis) for each bin of reader types.[24] The averages include the product of article features and the relevant article characteristic coefficients, plus beat × week FEs (for in-house beats) and the reader-type intercept for each user type. Grouping articles by beat in this way takes account of the correlations between article features that exist in the data, and gives a sense of which beats are most productive in generating output that induces readers to visit or subscribe. Component-level estimates of the article-characteristic multipliers β^V_x, β^R_x, with

[23] For computational efficiency, we compress our subscription opportunity observations into cells with the same menu, bin, day, and article (and record the number of observations and the share of subscriptions purchased). As such, we draw two sets of Dirichlet weights for each bootstrap iteration: for visit observations, we draw weights from Dirichlet$(1, \ldots, 1)$, treating the (article × day × bin) as the unit of observation. For subscription opportunity observations, we draw weights for each cell according to a Dirichlet distribution whose parameters are equal to the number of subscription opportunities observed at the given menu, bin, day, and article in the cell.

[24] Note that the x and y scales are not comparable to each other; one is the argument of a Poisson and the other is the argument of a logit.

Figure 6: Per-beat averages of visit and subscription utility, across all articles published in the data, given the estimates of our demand model parameters. The red points correspond to in-house authors, and the blue dot corresponds to articles sourced from wire services. For in-house beats, the utility includes the relevant beat × week fixed effects.



bootstrapped confidence intervals, are provided in Appendix B.2, Figures B.5 and B.6.

Unsurprisingly given the subscription rates, Bin 1 subscription utilities are uniformly much lower than the other two reader types. For the higher-propensity bins, however, hard news beats like Business, Health and Local News (which produce a disproportionate share of the newspaper's CIN and investigative content, see Figure A.5a) generally outperform the soft news beats like Entertainment and Sports. Almost all in-house beats outperform wire-sourced articles on both dimensions for Bins 2 and 3. For Bin 1, wire-sourced articles are at the bottom in traffic generation but average in subscription utility.

Figure 6 focuses on the article-specific component of reader utility that readers evaluate when a paywall blocks them from reading a specific article. Readers also get additional (discounted) value from access to everything the newspaper publishes over the term of their subscription, net of any resubscription fees they pay once the intro term elapses. Table 5 compares the full-bundle with the article-specific value for each reader type. The numbers in the table are dollar-denominated average values, aggregated over the full set of subscription options presented, using estimated price disutilities and article and bundle parameters. We average over all visitors to the newspaper, including the large majority who do not choose to subscribe. Full estimates of the bundle and paywall parameters $\alpha_x, b_x, \delta_x, \eta_x^R$ and κ_x^R, with bootstrapped confidence intervals, are provided in Appendix B.2, Figures B.7:B.8.

For Bin 1 readers, the subscription term offers essentially no additional value. For these readers, a hypothetical pay-per-view option (which would provide utility given by equation 5 with $B(o, d, x) = 0$) would be nearly equivalent in value to the full-access-for-an-intro-period structure that the newspaper actually offers.

For readers in Bins 2 and 3, however, the all-inclusive access accounts for most (roughly 95%) of the

value of the subscription. The levels of value are also several orders of magnitude higher than for Bin 1, corresponding to these readers' much higher average subscription propensity.

Table 5: Inclusive value decomposition across subscription structures, by user type.

Bin	Full Bundle	PPV
1	0.000280	0.000311
2	0.025114	0.001428
3	0.039847	0.001834

Notes: Values are in dollars, averaged over all visitors presented with a subscription option. "Full Bundle" is the standard intro term at a promotional price, during which the reader has full access to all articles published by the paper, followed by the option to renew monthly at a higher price. "PPV" or pay-per-view is a hypothetical subscription (not actually offered by the newspaper) that gives access to one article only, with no all-access term or renewal option.

While these exercises provide useful summaries of the demand model estimates, they take article production as given, and so they do not directly inform the newspaper's production incentives under alternative revenue models. In order to study this question, we move to explicitly modeling the newspaper's production process and its choices of product attributes under different objectives.

4 Supply Model

In this section, we present a stylized model of the supply of articles given slow-moving allocations of journalists across different beats and short-run shocks to the newsworthiness of different topics covered by these beats. Our model captures two key features of article production in our data: (1) higher levels of staffing in a beat correspond to more articles produced in that beat; and (2) the quality of the marginal article produced decreases as the number of articles in a beat increases. Our model makes predictions about the set of articles that would be published if the newspaper were to maximize expected readership, expected subscriptions, or a mixture of both. In this way, our model allows us to explore how different revenue sources might induce different allocations of journalists to beats and articles to production.

Diminishing marginal returns Allocating more staff to a beat will generally increase the number of articles that the newspaper produces in that beat. However, given a fixed set of newsworthy stories, each additional story may be less compelling than the previous ones. For example, if the newspaper staffs one reporter covering the local professional basketball team, that reporter can report on game results. If another basketball reporter is added, the second can additionally cover front-office moves and player trades. If more basketball reporters are added, the third, fourth and fifth will have to search increasingly further afield from the core of reader interest—the narrative of the team's performance through the season—to generate additional content.

In Table 6, we provide some suggestive evidence that the returns to allocating more staff to a beat are

diminishing in our data. To do so, we estimate the relationship between the *minimum* quality[25] of the articles published in a given beat each day, and the number of staff allocated to that beat on the same day. We expect, and find, a negative relationship: more staff assigned to a beat corresponds to more articles, but these additional articles generate lower reader interest on the margin. This relationship holds within beat and within month.

Table 6: Regression of minimum article quality on the number of reporters assigned to a beat.

	Min. Quality in Beat	
	(1)	(2)
Active Reporters	-0.0165***	-0.0223***
	(0.0051)	(0.0038)
Beat FEs	Y	Y
Month FEs		Y
Observations	8,649	8,649
R^2	0.25611	0.55040
Mean dependent variable	7.5806	7.5806

Notes: An observation is a beat × day. Article quality is the predicted log total visits to the article, based on the model reported in Figure B.1. The dependent variable is the minimum article quality among all articles published in that beat that day. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A model of article production given staffing We propose a parsimonious model of news article production given the levels of staffing in each beat and exogenous shocks to that beat's newsworthiness. On each day d, the newspaper observes a topic shock $z_{b,d}$ to topic beat b, a generic shock to interest in the newspaper w_d, and a monthly global trend shock $\xi^q_{m(d)}$.

We assume that the newspaper is able to produce a maximum of $\bar{N}_{b,d}$ articles on any given day. For each prospective article $j \in \{1, \ldots, \bar{N}_{b,d}\}$, the newspaper observes a *quality score* $q_{b,d,j}$ that captures the vertical value of article j with respect to the newspaper's objective (either readership or subscription propensity).

The average quality of articles increases in the interest shocks $z_{b,d}$ and w_d, but there remains substantial heterogeneity in article draws. To capture this, we assume that each prospective article quality score is drawn IID according to:

$$q_{b,d,j} \sim \mathcal{N}(\mu_{b,d}, \sigma_b) \text{ where } \mu_{b,d} = \alpha^q_b + \beta^z_q z_{b,d} + \beta^w_q w_d + \xi^q_{m(d)}. \tag{7}$$

Although the newspaper observes the prospective scores for the full set of $\bar{N}_{b,d}$ articles that it *could* theoretically write in a given beat, it is practically constrained by the number of journalists that are staffed in each beat. Given a level of staffing $A_{b,d}$, the newspaper produces a subset $N_{b,d} \leq \bar{N}_{b,d}$ of articles. We assume that staffing is fixed in the short term, so that there cannot be re-allocation of

[25] We measure quality as the predicted log total visits to the article, based on the model reported in Figure B.1.

journalists to accommodate particularly newsworthy days. Given an allocation of staff across beats $\{A_{b,d}\}$, the number of articles in each beat is determined independently. Specifically, we assume that $N_{b,d}$ is drawn according to an IID binomial distribution:

$$N_{b,d} \sim \text{Binomial}\big(\bar{N}_{b,d}; p_{b,d}\big) \quad \text{where } p_{b,d} = \text{logit}^{-1}\Big(\alpha_b^p + \gamma_b^p \log(A_{b,d})\Big). \tag{8}$$

The quantity equation (8) allows the productivity of staffing to vary by beat; it may be that the marginal entertainment writer can produce articles at a faster clip than the marginal politics writer, for example.[26] The log specification implies that in the case $A_{b,d} = 0$, $p_{b,d} = 0$; that is, the newspaper must assign positive staff to a beat to produce any articles in that beat.

In order to choose *which* subset of articles to publish, the newspaper ranks the articles in each beat in descending order of quality score $q_{b,d,j}$ and publishes the top $N_{b,d}$ articles. Note that because the quality score depends on the objective (traffic versus subscription revenue) the ranking of articles to publish will differ by objective.

It is worth briefly commenting on how this model relates to our stylized facts. First, by allowing beat-specific intercepts, variance terms and interest shifters, the model is able to capture substantial heterogeneity in the production patterns of different beats. Second, by endogenizing the number of articles that are produced as a stochastic function of beat staffing, the model is able to capture the effect of staffing allocation decisions on article production while still allowing for enough flexibility to match day to day variation in the data. Finally, the "top draws" model of article quality determination implies that holding fixed the underlying distribution of scores on a given day, the mth article published in a beat is distributed according to the mth order statistic of the underlying score distribution. The model captures the diminishing returns property because the mth order statistic is decreasing in m.

A model of optimal news beat staffing At the beginning of a given period, the newspaper decides how many journalists to hire for each beat. Because we cannot observe non-journalist staffing decisions or other budgetary considerations, we assume that the newspaper has a fixed annual budget for journalists in each year based on the observed numbers of journalists in our data. We model the newspaper's problem as one of constrained maximization given a fixed staff budget. The newspaper chooses the allocation of staff across beats to maximize a revenue objective subject to the constraint that the total number of journalists hired is constant.[27] We treat outsourced articles (those originating from wire services rather than written in-house) as constant throughout and do not vary them across counterfactual scenarios.

We consider two revenue objectives: (1) revenue from advertising, which we assume to be increasing in the number of visits, and (2) revenue from subscriptions. To model revenue from advertising, we

[26] See Figure A.6 for the differences in the data in average productivity across beats.

[27] In reality journalists in different beats may have different salaries, and in the next section we present some estimates of salary differences across beats. In our main counterfactual analysis, however, we treat all journalists as equally costly in order to focus on incentives due to reader demand rather than the supply of journalistic expertise.

assume that the newspaper receives a constant price per 1,000 pageviews ("cost per mille", or CPM, in industry terminology).[28] For a given candidate selection of staffing levels, we simulate the articles that would be produced based on our model of article production and sum the predicted total visits to all articles multiplied by the CPM rate. To model revenue from subscriptions, we simulate the articles that would be produced based on our model of article production, simulate visits to each article based on our visit model in Section 3, and randomly assign menus of offers to each simulated visitor based on their empirical frequencies. We then simulate subscription decisions based on our subscription model in Section 3 and sum the initial offer prices. We predict renewal revenue from each subscriber based on the empirical frequency of renewal for each offer we observe in the data, for each month following the end of the intro term.[29]

4.1 Estimation

We estimate our supply model with two specifications of quality scores matching our revenue objectives: (1) predicted log visits and (2) predicted log subscriptions. In each case, the predicted scores that the newspaper uses to rank articles are taken from our demand model estimates in Section 3 as follows.

Predicted log visits for article a take the form:

$$\widehat{q}_a^V = \log\left(\sum_{x=1}^{3}\exp\left(\frac{1}{||c(a)||}\sum_{c\in c(a)}\widehat{\zeta}^V_{c,w(d)} + \widehat{\beta}^V_x\theta_a\right)\right).$$

where $\widehat{\zeta}^V_{c,w(d)}$ are the estimated beat-by-week fixed effects at the week of article a's publication, for beats containing an author of article a. $\widehat{\beta}^V_x$ is the vector of estimated visit coefficients for article characteristics for reader-type x and θ_a is the vector of article characteristics for article a.

Predicted log subscriptions have the form:

$$\widehat{q}_a^R = \log\left(\sum_{x=1}^{3}\exp\left(\left(\widehat{\beta}^R_x + \widehat{\beta}^V_x\right)\theta_a + \hat{\eta}^R_x + \frac{1}{||c(a)||}\sum_{c\in c(a)}\widehat{\zeta}^R_{c,w(d)} + \widehat{\zeta}^V_{c,w(d)}\right)\right),$$

where $\widehat{\zeta}^R_{c,w(d)}$ are the estimated beat-by-week subscription fixed effects at the week of article a's publication for beats containing an author of article a. $\widehat{\beta}^R_x$ is the vector of estimated subscription coefficients for article characteristics for reader-type x, and θ_a is the same vector of article characteristics as before. $\hat{\eta}^R_x$ are the reader-type specific paywall subscription intercepts; we include these because they

[28] The constant price assumption implies that maximizing ad revenue alone is equivalent to maximizing visits. If the newspaper were to maximize a combination of ad revenue and subscriptions, the specific value of CPM would affect the weight that the newspaper places on visits compared to subscriptions. We do not have access to ad price data, and hence we focus on the polar cases of all weight on subscriptions versus all weight on visits.

[29] Renewal rates enter as $\hat{\pi}_{o,t}$ in Equation 4. Estimates of renewal rates in the data are reported in Figure B.3.

are part of the article-specific value. Note that the visit quality $\widehat{q}_a^V$ enters additively in the expression for subscription quality $\widehat{q}_a^R$ because subscriptions are the product of visits times subscription rates, and hence log subscriptions are the sum of log visits and log subscription rates.

In addition, we use Google Trends data for z_{bd} and w_d. Specifically, we use weekly Google Trends scores for our newspaper's name in its home media market for w_d. For z_{bd}, we generate query terms by sampling up to 25 articles per month from each beat, and passing the sampled articles' titles to the open-weight LLMs Mistral Small 3.2 and Qwen3 with instructions to generate ten search queries that "users might have used to find articles like these." We consolidated these terms, and collected weekly Google trends scores for each within the newspaper's media market. To reduce the dimensionality of our model, we standardize the scores for each query, and then average the topic-specific z-scores in each beat-week. We model monthly quality trends $\xi^q_{m(d)}$ with a month fixed effect.

The maximum number of articles that could have been published, $\bar{N}_{b,d}$, is not well-identified from the article data alone, as we observe only the number that actually were published, $N_{b,d}$. The binomial mean is $p_{b,d}\bar{N}_{b,d}$, and so these two parameters have equivalent effects on the mean prediction for $N_{b,d}$.[30] We therefore calibrate $\bar{N}_{b,d}$, setting $\bar{N}_{b,d} = \gamma^{\bar{N}} \log(1 + A_{b,d})$, where $\gamma^{\bar{N}}$ is the minimum value such that $\bar{N}_{b,d} \geq N_{b,d}$ for all days d and beats b.[31]

The remaining parameters of the article production model are estimated in the same way as our demand model, using maximum likelihood with inference via a Bayesian bootstrap procedure. To ensure internal consistency when solving for optimal beat staffing under different revenue objectives we use the supply model estimated with respect to the matching quality score specification (e.g. using visit scores for article production when allocating staff to maximize advertising revenue).

Because the supply model collapses articles to a one-dimensional quality score, we apply an additional step to generate the full set of article features for articles produced in counterfactual simulations. These features θ_a are needed as inputs to the demand model to predict visits and subscriptions. To generate article features from a predicted quality score, we sample a real article (in the data) from the same beat with a similar quality score, and use the sampled article's features.[32] In simulations, we draw a collection of potential article scores $\bar{N}_{b,d}$, as well as the number of articles that can be produced $N_{b,d}$, from the estimated supply model for each day d and simulated staffing in beat b. We then match the score of each article produced to a real article as described above. Finally, we predict visits and subscriptions based on the simulated articles and the parameter estimates from our demand model.

[30] They have different effects on the *variance* of $N_{b,d}$, as well as on moments of the published quality distribution, but in simulations we find that these two components are not easily separated in samples of the size we have in the real data.

[31] $\gamma^{\bar{N}} \approx 12.71$ in our data. See Figure B.9 for a visualization of the underlying data and the calibrated slope.

[32] Specifically, for each simulated quality draw $\tilde{q}$ for beat b, we draw the 10 real articles from b with quality scores q_a closest to $\tilde{q}$, and sample one from this set with probability weights equal to the normal density with mean $\tilde{q}$ and standard deviation 0.1 evaluated at q_a. This kernel sampling method approximates choosing the closest real article to $\tilde{q}$ but avoids repeatedly sampling the same article many times if multiple simulated draws fall into the same gap between real articles.

4.2 Results

We report parameter estimates and confidence intervals in Figures B.10-B.11. Using our supply parameter estimates, we simulate the consequences for article production, and subsequently for simulated visits and subscriptions, of different staff allocations.

Marginal impact of staff choices. In Figure 7, we show the estimated effects on traffic (top panel) and subscriptions (bottom panel) of adding or subtracting authors from each of the beats.

In each panel, we assume the newspaper selects articles to publish according to the corresponding objective. In the visit panel, the newspaper ranks articles according to predicted visits ($\hat{q}_a^V$); in the subscription panel it ranks articles according to predicted subscriptions ($\hat{q}_a^R$). In both cases, the newspaper then publishes the top $N_{b,d}$ articles (according to the relevant ranking) for each beat-day given available staffing.

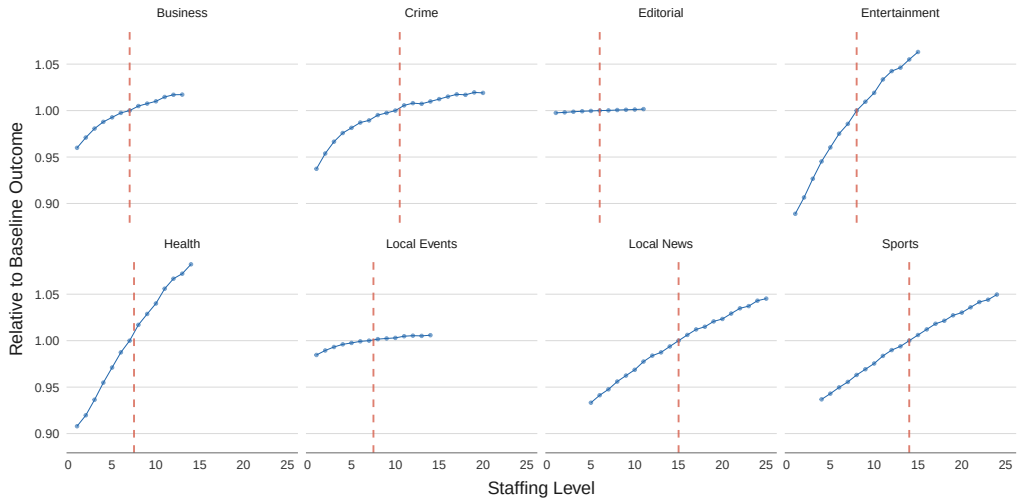
This exercise illustrates the differences across beats. Adding Entertainment or Crime writers increases the newspaper's traffic but has much weaker effects on subscriptions. Adding Local News writers has a weaker effect on traffic but a stronger effect on subscriptions than adding Entertainment writers. These comparisons all take into account differences across beats in author productivity (through the parameters α_b^p, γ_b^p), as well as differences in the mean and variance of article quality and responsiveness to interest shocks (through the parameters $\mu_{b,d}, \sigma_b$).

Optimal staffing. Next, we use a stochastic optimization method to find the optimal staffing choice that satisfies the newspaper's budget constraint: the visit- or subscription-revenue-maximizing allocation that holds fixed the total number of journalists that the newspaper employs.[33] In effect, this exercise assumes that the newspaper faces a staff budget constraint, and that journalist salaries are constant across beats. This assumption isolates demand- rather than cost-driven changes in optimal staff allocation under different revenue objectives.

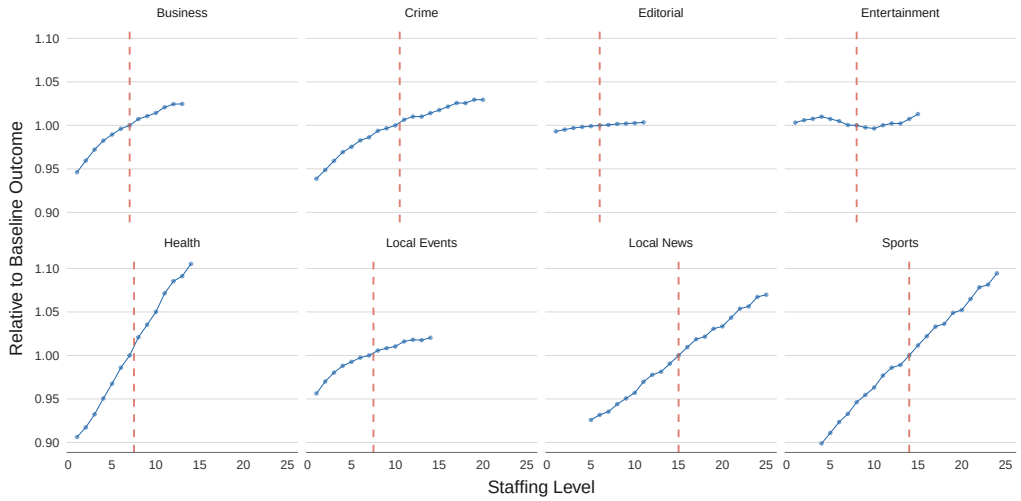
Figure 8 visualizes the changes in optimal staffing allocated to each beat under different objectives—optimizing ad revenue versus optimizing subscription revenue—relative to the baseline allocation that we observe in the data.

We find that there are beats which are understaffed (Health) or overstaffed (Crime) regardless of objective. Note that we are not accounting for salary differences here, and these consistent differences from maximizing behavior may be due to differences in the market wages of crime reporters versus health experts, for instance.

[33] Because the objective function is costly to evaluate, we use "surrogate" optimization: we sample allocations from the feasible set, fit a flexible polynomial to the sampled points, then optimize the surrogate rather than the original function. We then evaluate the actual objective at the surrogate optimal point, add it and neighboring points to the set of evaluated points, re-fit the surrogate, and repeat until convergence.



(a) Visits



(b) Subscriptions

Figure 7: Marginal effects of adding or subtracting additional staff on visits (top panel) and subscriptions (bottom panel). The y-axis is total visits (top) or subscriptions (bottom), relative to the baseline level of the outcome. We consider a range of ± 10 *staff compared to the baseline for the beat (bounded below at 0).*

For the remaining beats, the direction of staff change depends on the objective. Maximizing traffic leads to cuts at the Local News beat (the beat with the largest emphasis on political news), Local Events, Sports and Business, paired with a large expansion of the Entertainment beat. Under the subscription-optimal allocation, the picture is reversed, with Local News, Local Events, and Sports growing at the expense of Entertainment. It is apparent that the two objectives produce quite different allocations of writers across the beats, and a different composition of news produced for readers. Overall, the ad-driven model appears to lead to cuts at the more socially valuable sections of the paper, while the subscription-driven objective preserves or expands these sections.

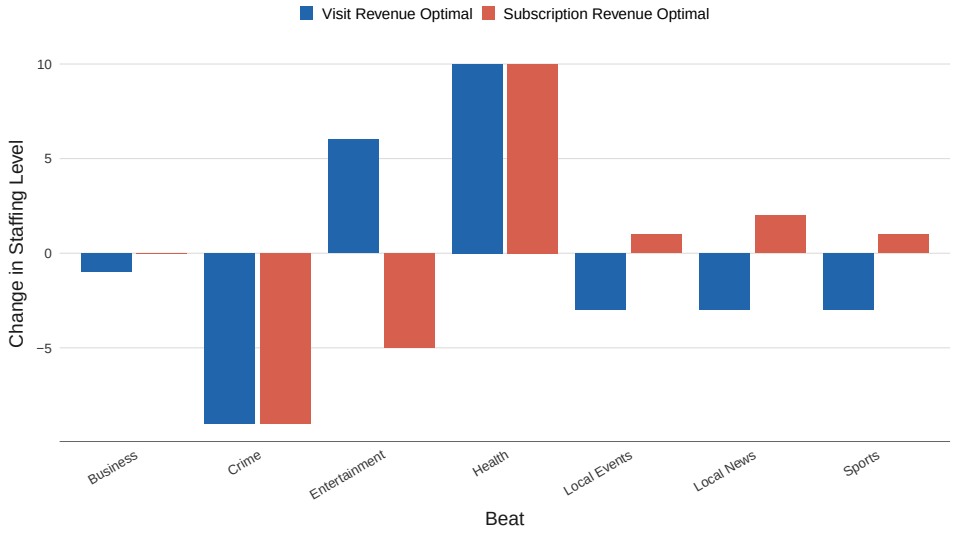


Figure 8: Optimal Staffing Under Different Objectives

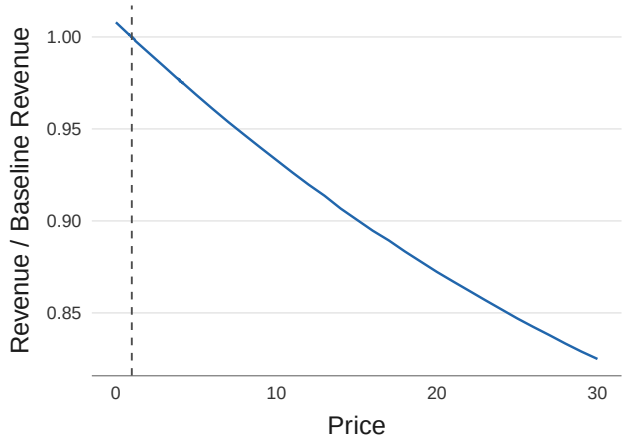


Figure 9: Total subscription revenue (relative to baseline) as a function of intro price of the standard offer

Optimal pricing. We conduct two optimization exercises to compute the optimal pricing from the newspaper's perspective. In the first, we vary the introductory price of the newspaper's most common subscription offer, which has a three-month intro term followed by renewal at $14 / month unless the reader decides to cancel. We search for the introductory price that maximizes the newspaper's total revenue (including both initial and expected resubscription revenue). Figure 9 shows that resubscription revenue dominates, such that the optimal (subscription-revenue-maximizing) introductory price for this offer is $0. The observed price is very low ($0.99), suggesting that the newspaper realizes this, and charges a nominal introductory price not for the revenue it generates but simply to store credit card information for future renewals.

Pay-per-view. Our second pricing exercise considers an alternative subscription model: "pay-per-view." In this version, we set the intro term and the probability of resubscription to zero, which in our

demand model corresponds to setting the bundle value $B(o, d, x) = 0$ in readers' utility. This exercise models an alternative product where rather than offering all-access subscription, the newspaper sells articles a la carte. The revenue-maximizing price here ($21) is much higher than the observed intro prices in our data but less than the average total price (intro plus renewal(s)) we observe users paying given the observed resubscription rates and resubscription prices. Because resubscription revenue is such an important component of revenue, there is *no* pay-per-view price that generates equivalent revenue for the newspaper; even at the optimum price the revenue from pay-per-view is only about 61% of that generated by the observed three-month intro / renewal combination.

Marginal revenue and production subsidies. Finally, we consider which beats have the highest return to investment in staff, both in subscription revenue and in the production of socially valuable news content. While exact journalist salaries are heterogeneous and not available to us, we are able to approximate average wages for different types of journalists using salary data from the vendor Revelio.[34] We computed the average modeled salary (using Revelio's proprietary salary model) for journalists in the newspaper's metro area specializing in Arts and Entertainment, Business, Politics, Sports, Science and Environment, and Health. We filter to journalists using Revelio's job classifications, and assigned specialties based on keyword searches of job descriptions. Politics and business writers are the most expensive, earning about 30% more than sportswriters and entertainment writers.

We then compute the marginal subscription revenue (the change in subscription revenue when adding one additional staff to a beat) and compare to the estimated salary cost from Revelio, over the same range of staff allocations considered before. Figure 10 plots the results of this exercise. In spite of the relatively higher salaries earned by politics and business writers compared to entertainment or crime writers, Local News and Business are closer to break-even at the observed allocation. Health reporters generate the highest fraction of their average salary in marginal subscription revenue.

However, *all* beats are underwater at the observed allocation and at all tested alternative points. Although the more socially valuable parts of the paper generally also produce more private value, the fraction of this private value that can be captured by the newspaper through digital subscriptions is insufficient to cover its marginal salary costs.

Given this result, we move to estimating the subsidy—from government or philanthropic sources—that would be required to maintain or expand levels of news production. As a measure of donor objective, we consider the newspaper's production of investigative articles, a category of journalistic output with positive social value but questionable economics for news producers (Hamilton, 2016). We compute changes in the production of investigative articles as we increase staff in each of the sections, and compare to the subsidy that would be required to offset the newspaper's net loss from the hire (which can be zero, if the estimated net gain in revenue from the hire is positive). This allows us to measure the returns to a public investment, e.g. through a payroll subsidy, in supporting news organizations.

[34] Revelio's data is based on public LinkedIn profiles, supplemented with its own data and models of salaries and job functions.

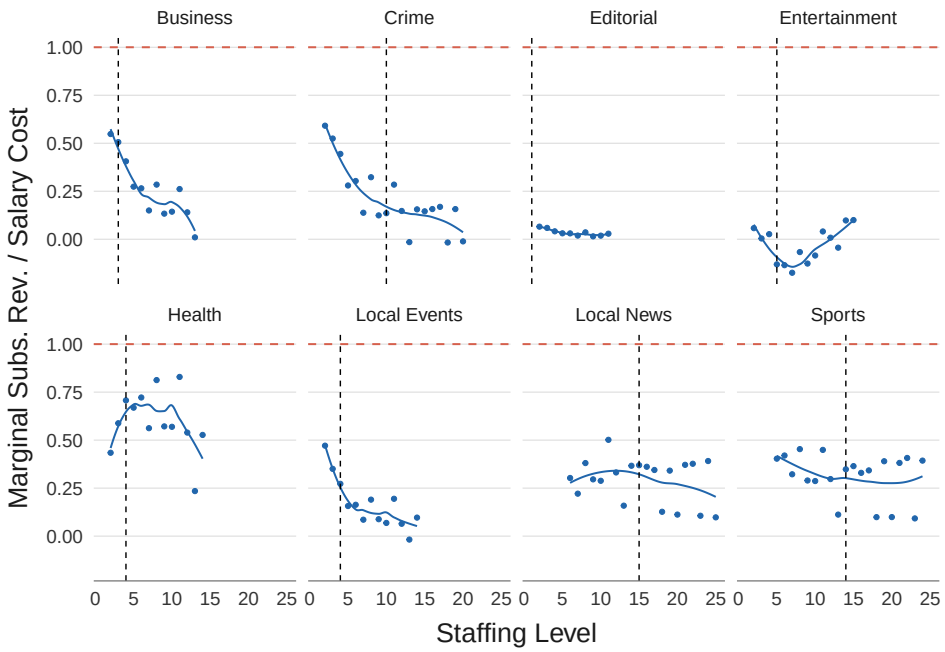


Figure 10: Estimated ratio of marginal subscription revenue to journalist salary, by beat. Lines are local linear smoothers.

Figure 11 shows how the total number of investigative articles that the newspaper produces changes, as we add and subtract staff from different beats. The Local News and Health beats generally have steeper gradients between staffing and investigative article production, and so they outperform the other beats on this dimension.

Figure 12 combines the analysis of marginal production and marginal revenue losses together. To do this, we compute the subsidy needed to raise the marginal subscription revenue net of salary costs to zero for all points depicted in Figure 10. We compute marginal annual investigative article production by taking differences between adjacent points in Figure 11. We then compute the minimum subsidy needed to produce a given number of additional investigative articles per year without losing money for the paper, allowing only additions of new staff (i.e. we consider only points to the right of the dashed lines in Figure 10).

Our results indicate that the marginal subsidy needed to increase the production of investigative articles without inducing a net revenue loss is roughly $11,000. However, because each subsequent investigative article is harder to write, the cost of inducing additional articles increases. The cost of subsidizing the first 20-30 additional articles per year is relatively flat at around $25,000—in part because of the relatively low marginal losses from local increases in beats like Health and Local News. Beyond that, however, costs ramp up more quickly, reaching $50,000 per article per year at 35 additional articles per year and more than $100,000 by 55.

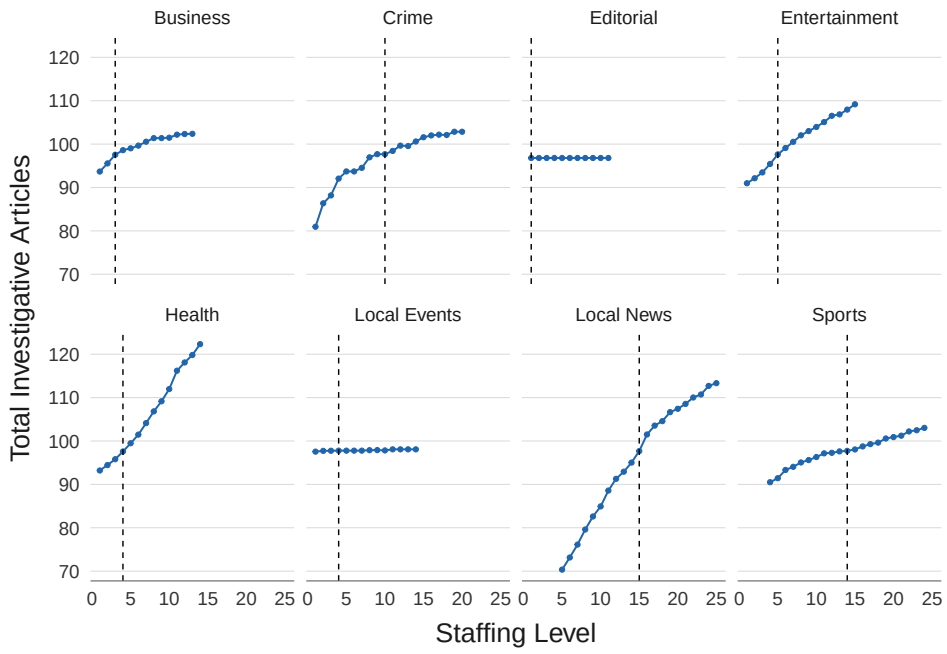


Figure 11: Changes in annual investigative article production by the newspaper, as staff are added to or removed from beats. The y-axis is the total number of investigative articles produced annually across all beats.

5 Conclusion

The proliferation of real-time metrics—and consequent visibility of editorial management into the kinds of news that are "fit to click"—has clearly influenced the practice of journalism in the twenty-first century (Petre, 2021). However, our results show that information about web traffic responses alone can be very misleading for understanding the relative values that readers place on different types of journalistic output. Because news readers' marginal willingness to pay in attention is systematically different from their marginal willingness to pay in dollars, click-based modes of evaluation substantially underestimate the private value of public-interest news content.

This difference is crucial to understanding the production incentives that media outlets face under different revenue models. We show that the staff of journalists that would maximize traffic is quite different from the staff that would maximize total subscription revenue. Sections of the newsroom that produce longer articles at a slower pace—which have suffered in the internet era—could see a revitalization if newspapers were to fully commit to a subscription-supported model and abandon the quest for clicks and advertising impressions.

Changes in advertising technology appear to be pushing outlets in this direction already. Large online platforms' troves of personal data have given them a massive advantage in consumer targeting precision over even the largest media outlets. In turn, such platforms have captured the lion's share of the digital advertising market and made revenue growth from digital advertising a difficult proposition for news producers. And new AI-enabled search and summarization technologies appear likely to

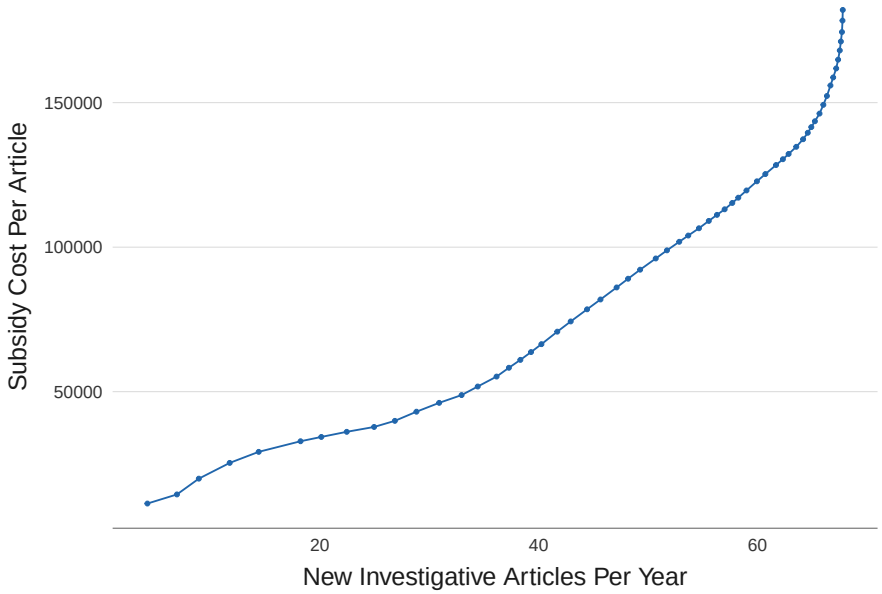


Figure 12: The subsidy per article required to add a given number of investigative articles each year above the newspaper's baseline production.

dramatically shrink traffic flowing to news sites. News outlets may have little choice but to focus on their subscription businesses out of necessity, given this competitive landscape. The journalistic beats which are the most commercially viable under the digital subscription model are also those which produce the most CIN and investigative content.

Optimism on this score should be tempered by the fact that our main counterfactual staffing exercise speaks only to the *relative* direction in which production incentives drive staff allocation, and are subject to the constraint that the overall staffing level remain fixed at the observed level. It may still be the case that reporters' marginal revenue product from digital subscriptions fails to cover their salary and overhead costs, and thus that the production of socially valuable information would be in jeopardy regardless of producers' business models.[35] Although our cost data is limited, we are able to show that digital subscriptions are not, on their own, likely to be a sustainable source of support for journalism. Production subsidies from governmental or philanthropic sources are needed to preserve the production of news in the current environment. However, the fact that socially valuable content is closest to commercial viability means that a donor interested in subsidizing greater production of investigative or accountability journalism could do so at relatively low cost, at least for small increases relative to the current baseline production.

We conclude that asking readers to pay for access to news content favors the production of political, economic, public-health and other "critical information needs" pieces more than does the sale of advertising against such content. Whether the returns to the subscription model are large enough to make news organizations sustainable in the long term without subsidies remains an open question.

[35] Our data sharing agreement with the newspaper prevents us from releasing absolute revenue figures.

References

Angelucci, C. and J. Cagé (2019): "Newspapers in times of low advertising revenues," *American Economic Journal: Microeconomics*, 11, 319–64.

Arceneaux, K. and M. Johnson (2013): *Changing Minds or Changing Channels?*, Chicago: University of Chicago Press.

Besley, T. and R. Burgess (2002): "The Political Economy of Government Responsiveness: Evidence from India," *Quarterly Journal of Economics*, 117, 1415–1451.

Blei, D. M. and J. D. Lafferty (2006): "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, Pittsburgh, Pennsylvania: ACM Press, 113–120.

Cagé, J., N. Hervé, and M.-L. Viaud (2020): "The production of information in an online world," *The Review of Economic Studies*, 87, 2126–2164, publisher: Oxford University Press.

Christin, A. (2020): *Metrics at Work*, Princeton University Press.

Djourelova, M., R. Durante, and G. J. Martin (2025): "The Impact of Online Competition on Local Newspapers: Evidence from the Introduction of Craigslist," *Review of Economic Studies*, 92, 1738–1772.

Durante, R., P. Pinotti, and A. Tesei (2019): "The political legacy of entertainment TV," *American Economic Review*, 109, 2497–2530.

Ewens, M., A. Gupta, and S. Howell (2023): "Local Journalism under Private Equity Ownership," .

Fan, Y. (2013): "Ownership Consolidation and Product Characteristics: A Study of the US Daily Newspaper Market," *American Economic Review*, 103, 1598–1628.

Friedland, L., P. Napoli, K. Ognyanova, C. Weil, and E. J. Wilson III (2012): "Review of the literature regarding critical information needs of the American public," Tech. rep.

Gentzkow, M., E. L. Glaeser, and C. Goldin (2007): "The Rise of the Fourth Estate: How Newspapers Became Informative and Why It Mattered," in *Corruption and Reform: Lessons from America's Economic History*, University of Chicago Press.

George, L. M. and J. Waldfogel (2006): "The New York Times and the market for local newspapers," *American Economic Review*, 96, 435–447.

Hamilton, J. T. (2004): *All the news that's fit to sell: How the market transforms information into news*, Princeton University Press.

——— (2016): *Democracy's Detectives: The Economics of Investigative Journalism*, Cambridge, Mass.: Harvard University Press.

Lippmann, W. (1922): *Public Opinion*, New York: Harcourt, Brace & Co.

Mahone, J., Q. Wang, P. Napoli, M. Weber, and K. McCullough (2019): "Who's Producing Local Journalism? Assessing Journalistic Output Across Different Outlet Types," Tech. rep.

Miller, K., N. Sahni, and A. Strulov-Shlain (2023): "Sophisticated Consumers with Inertia: Long-Term Implications from a Large-Scale Field Experiment," .

Nussbaum, Z., J. X. Morris, B. Duderstadt, and A. Mulyar (2024): "Nomic Embed: Training a Reproducible Long Context Text Embedder," _eprint: 2402.01613.

Petre, C. (2021): *All the News That's Fit to Click*, Princeton University Press.

Petrova, M. (2011): "Newspapers and parties: How advertising revenues created an independent press," *American Political Science Review*, 790–808.

Pew Research Center (2021): "Newspapers Fact Sheet," Tech. rep.

Prior, M. (2007): *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*, Cambridge University Press.

Snyder, J. M. and D. Strömberg (2010): "Press Coverage and Political Accountability," *Journal of Political Economy*, 118, 355–408.

Turkel, E., R. Owen, A. Saha, G. J. Martin, and S. Vasserman (2021): "Measuring Investigative Journalism in Local Newspapers," *Proceedings of the National Academy of Sciences*, Forthcoming.

Appendices

Appendix A Data construction details

A.1 Predicted article features

To determine whether each article in our sample falls into each of the eight CIN categories (Emergencies and Public Safety, Public Health, Education, Transportation, Environment and Planning, Economic Development, Civic Life, and Political Life) and six non-CIN categories (Business, Entertainment, Real Estate, Sports, Things to do, and Opinion Columns) we use the following supervised learning approach.

We first generated a set of low-level article features for each of the articles in our dataset. These include 768-dimensional embedding vectors generated from the pre-trained model of Nussbaum et al. (2024); a set of 25 consolidated section indicators, indicating that the article ran in e.g. the World News section;[36] counts of the number of times a local politician title (such as "mayor", "supervisor" and so on) appeared in the article; and the count of the number of times a place name of a city, county or town inside the paper's media market was mentioned in the article.

We next manually identified a set of positive examples of articles that fell into each of the 8 CIN and 6 non-CIN categories. We generated possible matches using keyword and section searches (for example, searching for articles in the Sports section to generate articles for the Sports category), and then manually validated each. All categories had at least 200 positive training examples identified by this method.

For each category, we estimated a separate L1-penalized logistic regression of an indicator for the article falling into the category on the features described above. All other articles in our training set were used as negative examples; for example, articles labeled as Sports-related were used as negative examples in fitting the model predicting Environment and Planning content.[37] We used 10-fold cross-validation to choose the penalty level for each of the L1-penalized logistic regressions.

Figure A.1 shows the correlation of the predicted features with each other and with the directly-measured local, in-house, and wire-service indicators. Of note is that the Emergency, Education, Economic Development, and Civics categories, as well as Real Estate, all positively correlate with the local-place-name indicator. Politics and Civics are also fairly correlated with each other, as are Economic Development and Business and Entertainment and Things to Do. Most other correlations

[36] To reduce the sparsity of the section features, we combined the 103 distinct section titles appearing in our articles dataset into 26 grouped sections. For example, we combined the separate sections for each of the area's local professional and college sports teams into a single Sports section.

[37] A small number of articles received positive labels in multiple categories; these were excluded from the set of negative examples for all categories into which they fell.

are small and negative.

Figure A.2 shows the time trends of the average feature weight in the Health (A.2a) and Politics (A.2b) categories. The COVID-19 pandemic is readily visible in spring to summer of 2020, as are the November elections in 2020 and 2022. In addition to the daily means, the figures plot the 90-day rolling means of the feature weight; the rolling mean of different article types over the intro term of a subscription (typically 90 days) will influence readers' perception of subscription value in our model.

A.2 User attributes

Figure A.3 shows the distribution of first-dimension principal component scores from PCA applied to a matrix consisting of the user's average reading depth of articles previously read, the total word count read by the user, the total number of previous articles visited, the word count of articles read in each of the previous six weeks, and the number of paywalls and registration walls encountered in each of the previous six weeks.

Figure A.4 shows the distribution of estimated probabilities of encountering a paywall, split by user bin (1, 2, or 3). These probabilities are estimated for each user-article visit from a logistic regression model where the outcome is a binary indicator for the user having encountered a paywall on that article. The model takes as inputs the same reading history variables that determine the PC1 score, as well as a number of attributes of the user-session such as: the user's browser and operating system; whether the referring website was a news aggregator, a social media platform, a search engine, or the newspaper's home page; and whether the user had ad-blocking software installed in their browser. The paywall probabilities are, unsurprisingly, generally higher for users in the higher bins, but within bin there is substantial variation in likelihood of hitting a paywall.

A.3 Author clustering

Our supply model (Section 4) groups authors into "beats" using the features of articles written by each author. We apply a k-means clustering algorithm to author-month observations. Each author-month observation is the average feature vector of all articles written by a given author in a given month; authors are therefore able to move from beat to beat over time. Figure A.5 shows the typical features of articles written by authors assigned to the cluster (panel A.5a) and the number of staff assigned to each cluster over time (panel A.5b).

Authors' typical productivity also varies across beats. Figure A.6 shows the average number of articles produced per month by each author associated with a given beat.

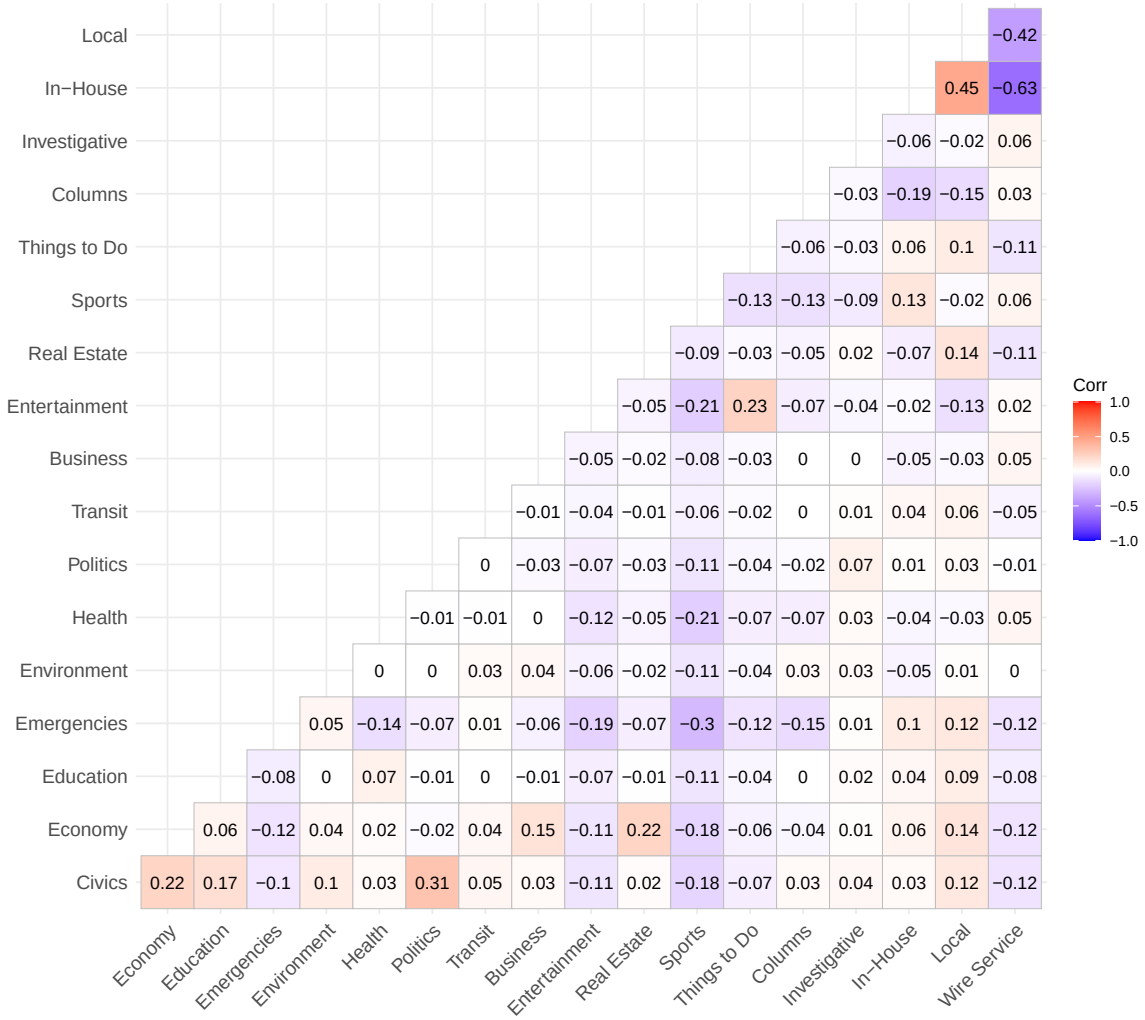
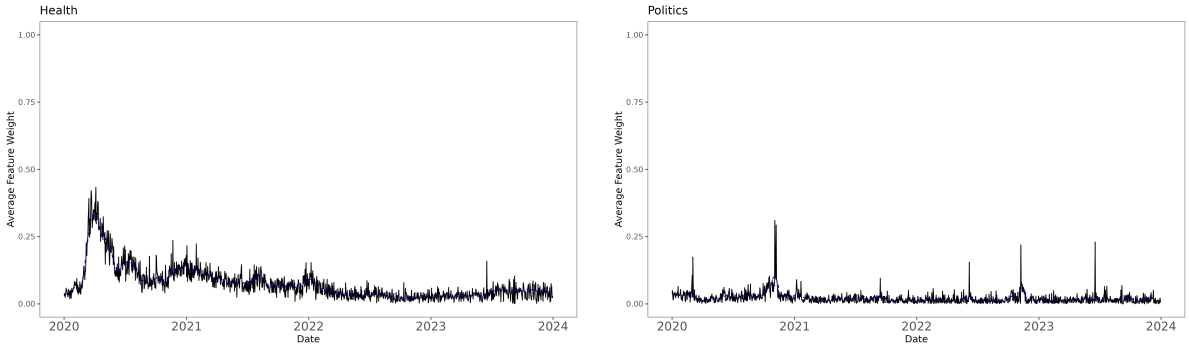


Figure A.1: Correlations between Article Features. "Local" means the article mentions at least one place name corresponding to a city or county contained in the newspaper's home media market. "In-House" means the article has at least one author who is a permanent employee of the newspaper; "Wire Service" means the article was sourced from the Associated Press, Reuters, or some other syndication service. All other features are content measures whose construction is described in Appendix section A.1.



(a) Health (b) Politics

Figure A.2: Time trends of the average feature weight in the Health and Politics categories. Lines show the average of the feature weight across all articles published on a given day.

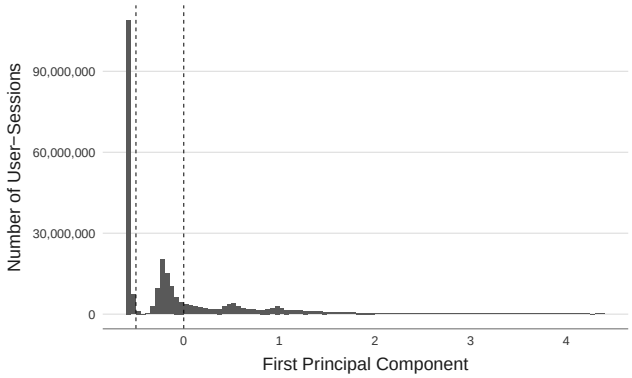


Figure A.3: Histogram of First-Dimension Principal Component scores (PC1). The score is derived from principal components analysis applied to a matrix consisting of a set of variables measuring the user's intensity of interaction with the site in the past six weeks, described in Section 1. PC1 varies by user-session and is used to bin users into discrete types. Dashed lines indicate the boundaries separating bins 1 from 2 and 2 from 3. To improve the visualization, the histogram is truncated at 5; there are an additional 7.76M user-sessions with PC1 greater than 5 not plotted.

Appendix B Estimation details

B.1 Additional reduced form traffic and subscription regressions

B.1.1 Visits

Figure B.1 shows the results of a set of regressions of the number of unique visitors to an article (inverse-hyperbolic-sine transformed to account for the right skew in the distribution) on the article attributes described in the data section. We interact each of our 14 categorical content variables (8 FCC-defined Critical Information Needs categories and 6 non-CIN categories) with an indicator for the article being written by a staff author.

Consistently for all user types and almost all categories, articles written in-house generate substantially more traffic than outsourced articles. The important exception is the Columns category; some of

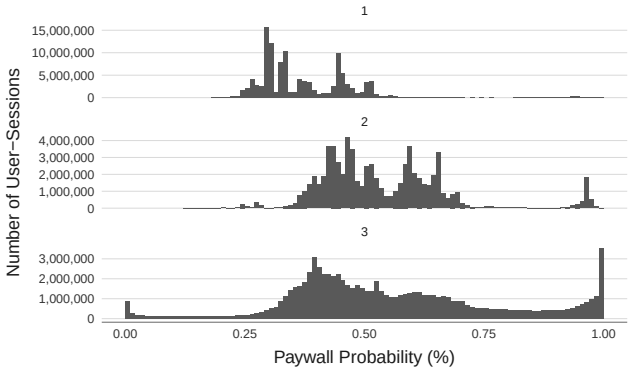
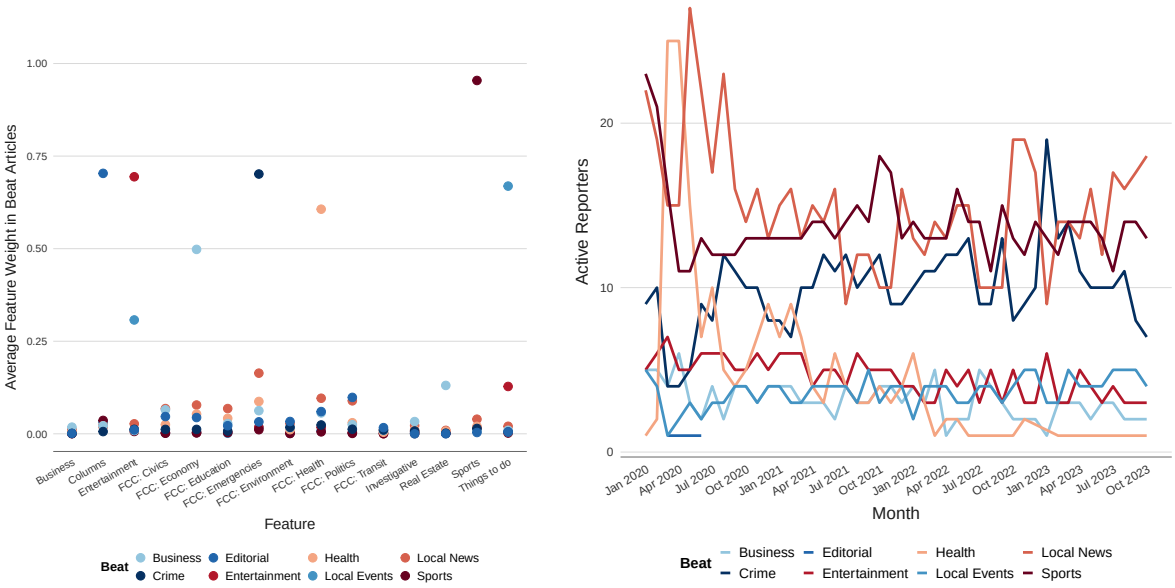


Figure A.4: Histogram of estimated propensities of encountering a paywall by user-article visit. The histogram is split by user type (1,2, or 3) defined by the user's PC1 score as described above.



(a) Average feature weight, by journalistic beat.

(b) Staffing levels, by journalistic beat.

Figure A.5: The result of our author clustering procedure. The left panel shows the average feature weight among all articles written by authors assigned to the beat. The right panel shows the number of authors assigned to each beat over time.

the site's biggest traffic drivers are syndicated advice columns. Health and Economic Development articles also appear to generate substantially more traffic than the other categories. Investigative articles written in-house generate modestly more traffic than articles with similar topical content but not categorized as investigative by the Turkel et al. (2021) method.

B.1.2 Subscriptions

We next turn to subscription decisions. We ask, conditional on being paywalled on an article, how much more likely is a non-subscriber to choose to subscribe if the article she was trying to read has

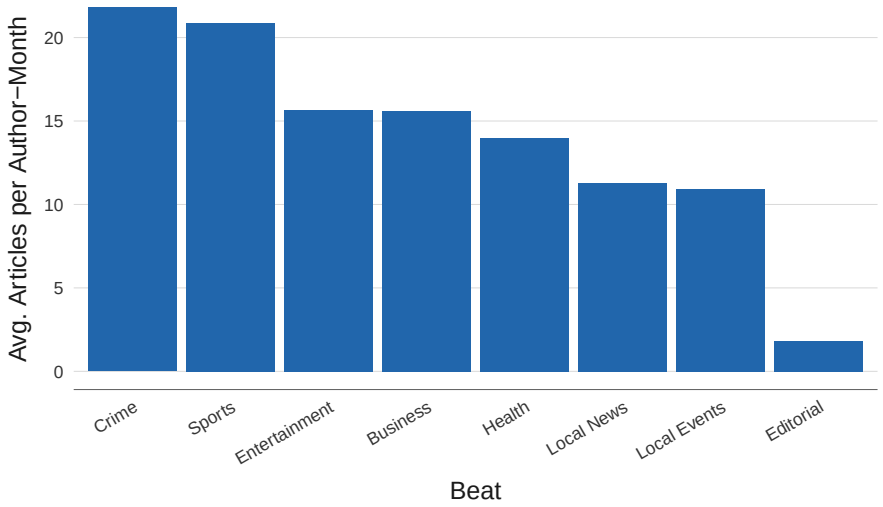


Figure A.6: Average articles published per author-month, by beat.

some particular feature, than if it does not? Again, the identifying assumption here will be that being paywalled is exogenous to unobserved article and user characteristics once we condition on the past reading history. Figure B.2 shows the results of a regression where the sample is not-currently-subscribed users who encounter a paywall when attempting to read an article. The dependent variable is an indicator for the user signing up for any paid subscription plan. Because the available offer pricing varies over time, we use date fixed effects to capture possible differences in subscription propensity deriving from lower-cost offers being available on some days and not others.

The results show that our binned indicator of user type reliably captures differences in taste for the newspaper. Our user bins are sequentially ordered in their responsiveness to article features.[38] The lowest bin (making up roughly half of all user-sessions) has no detectable response to any feature; this group almost never subscribes no matter where it encounters a paywall.

Comparing to the results for traffic, there is still generally a pattern that articles written in-house are more likely to convert readers to subscribers, especially those in the higher bins who have higher baseline subscription rates. Interestingly, the syndicated Columns category, which Figure B.1 shows is one of the biggest traffic generators, has a large negative effect on subscription propensity, particularly among the most-likely subscriber group. This is true even though all groups appear to desire to read articles of this kind. Similarly large negative coefficients are observed for outsourced Politics, Health, and Emergencies articles.

Comparing within the staff-written article types, the eight CIN categories defined by the FCC generally outperform "softer" news in terms of converting potential subscribers. Investigative pieces also increase subscription propensity, though again this effect is modest. At least among the type of users who are actually willing to pay for the product, wanting to read an article that is investigative, or locally

[38] They are also ordered in base subscription rates, though our data sharing agreement prevents us from reporting these.

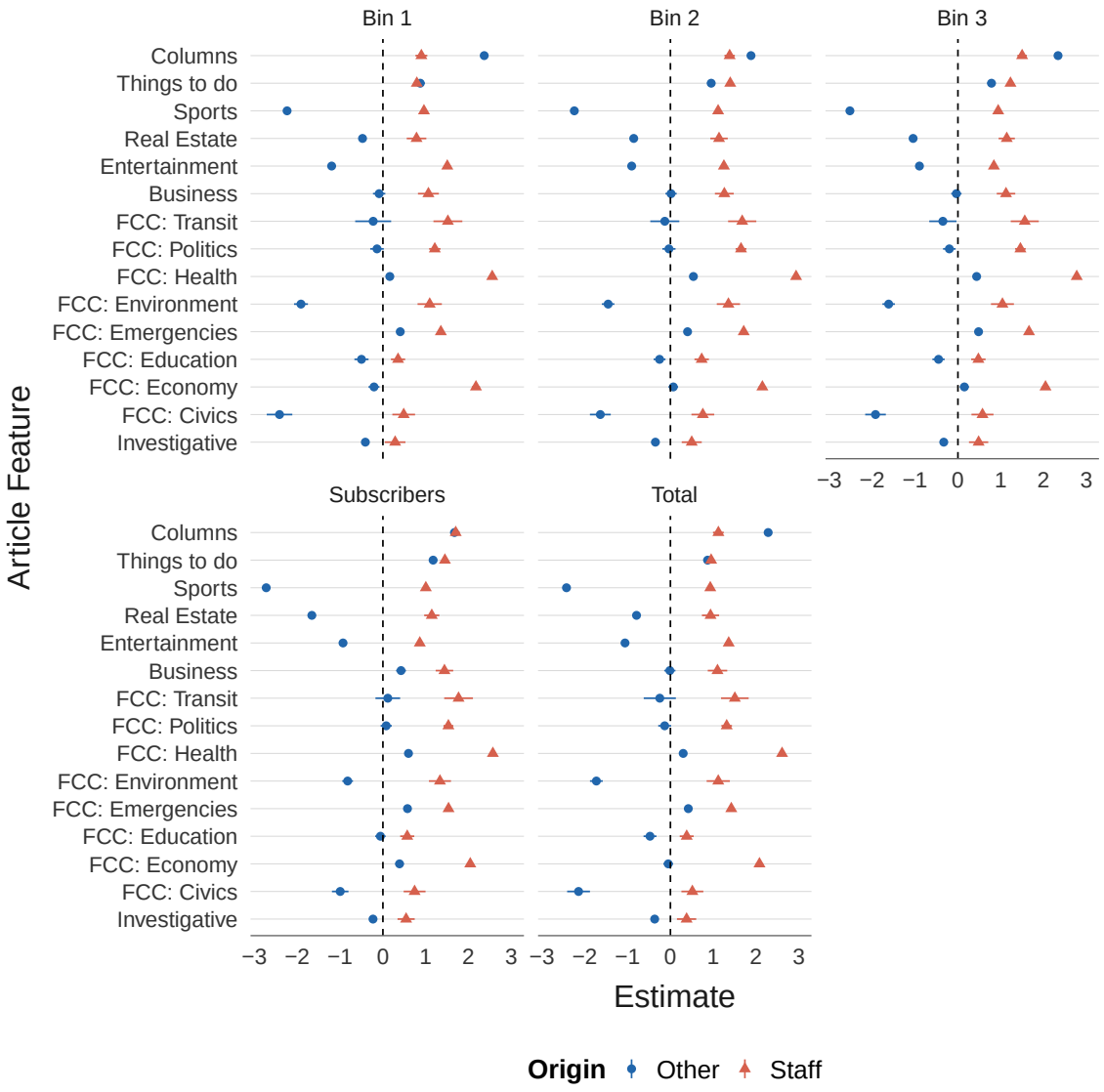


Figure B.1: Correlates of article traffic. Each panel shows the coefficient and its 95% confidence interval of the indicated feature in a model of IHS-transformed visits to the article within 14 days of publication. In the top three panels, the outcome is visits by each non-subscriber subtype; the bottom two estimate effects on visits by current subscribers and by all visitors, respectively. All models include fixed effects for date of publication and also cluster standard errors by date of publication.

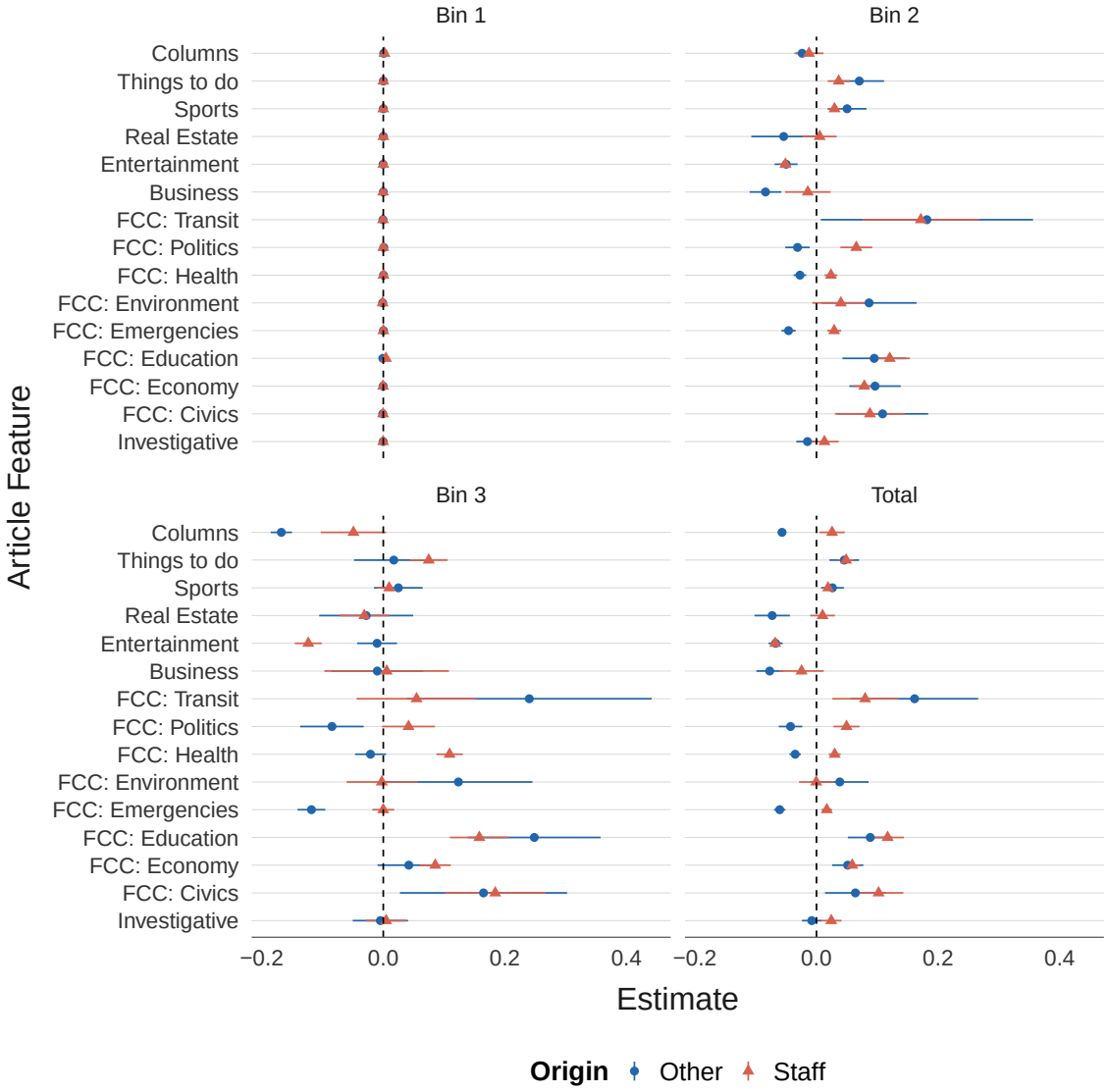


Figure B.2: Correlates of subscription, conditional on the features of the article on which a user encountered a paywall. Each panel shows the coefficient and its 95% confidence interval of the indicated feature in a linear model of subscription rate, conditional on hitting a paywall. An observation is a distinct combination of day, article viewed, user bin, and menu of subscription plans offered. Observations are weighted by the total number of user-sessions in that day-by-article-by-menu-by-bin cell. The dependent variable is the fraction of user-sessions in the cell which purchased any paid subscription plan. All models include fixed effects for visit date and also cluster standard errors by visit date.

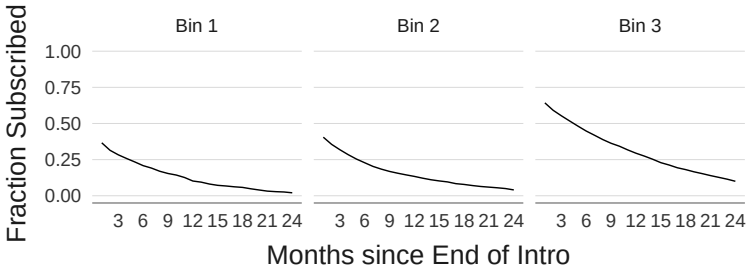


Figure B.3: Renewal rates by month. We plot, for each user type, the estimated conditional probability of remaining subscribed for a given number of months after the end of the intro term for the offer to which the user subscribed.

focused on an issue of public importance, is a motivator for subscribing to the newspaper.

Finally, we examine resubscription rates. Since the resubscription prices we observe in the data are much higher than the introductory price that applies during the intro term (often by a factor of 10 or more), the total subscription revenue that the newspaper earns is highly dependent on how likely it is that consumers renew their subscriptions beyond the end of the intro term. Figure B.3 plots, by user bin, the fraction of users who subscribed to any offer, who remain subscribed for 1, 2, 3, ... 12 months following the end of the intro term of their offer.[39] There is a clear pattern of over-time decay, but for all user types a significant fraction of users remain subscribed well beyond the end of the intro term.

B.1.3 Bundle value

The preceding analysis of subscriptions considers only the attributes of the article on which the user is paywalled. Our demand model, however, also contains a forward-looking "bundle value" which represents the discounted present value of the articles the user expects to read in the future, over the term of the subscription. The feature of the data that we seek to capture with this element of the model is that readers often subscribe to a longer-term subscription, even though shorter-term subscriptions are available and have lower upfront costs. If readers are purely myopic and think only about the value of the current article they are trying to read, the long-term and the short-term subscription are the same good — both allow reading the current article — and thus such choices cannot be rationalized unless readers prefer to pay higher prices for the same good.

This pattern can be seen in Table B.1. When we account only for the features of the article which the user visited, we estimate a positive coefficient on the offer price on the likelihood of subscription. This is because longer-duration offers are more expensive, and the price measure in Table B.1 is in absolute terms, not scaled by the length of the offer. In the remaining columns, we add the rolling average of each feature across all articles published in a given rolling window length around the visit time. We

[39] We cannot directly observe if a user remains subscribed, as we do not have direct access into the newspaper's billing system. Instead what we can measure is whether or not the user returns to the newspaper's website and is identified as a current subscriber by the paywall system. As it is possible that users remain subscribed but do not visit the website, the numbers in Figure B.3 are a lower bound.

multiply these rolling averages by the length of the subscription offer term. These offer-length-scaled rolling average features capture the value of reading over the term of the subscription, if the user's forecast of future articles was a constant equal to the average of articles actually published in a given window length around his or her subscription date.[40]

Table B.1: Estimates of price sensitivity to promotional price.

Window Length	None	15d	30d	45d	60d	90d
Intro Price	$1.37 \times 10^{-5**}$ (6.9×10^{-6})	$-1.97 \times 10^{-6**}$ (7.63×10^{-7})	-1.37×10^{-6} (8.51×10^{-7})	$-1.6 \times 10^{-6*}$ (8.13×10^{-7})	$-1.68 \times 10^{-6**}$ (8.03×10^{-7})	$-1.45 \times 10^{-6*}$ (8.13×10^{-7})
Article Features?	Y	Y	Y	Y	Y	Y
Rolling-avg. Features?	N	Y	Y	Y	Y	Y
Year FEs	Y	Y	Y	Y	Y	Y
Day-of-week FEs	Y	Y	Y	Y	Y	Y
Observations	13,485,953	13,605,131	13,510,895	13,390,457	13,238,094	13,007,504
R^2	0.0003	0.0007	0.0007	0.0007	0.0007	0.0007
Mean dependent variable	0.004	0.006	0.006	0.006	0.006	0.006

Notes: An observation is a unique combination of user type, date, article visited, offer menu displayed, and paywall status. Observations are weighted by the number of users in the cell. The left-hand side variable is an indicator for subscribing to any offer. In the first column only features of the visited article are included; in the remaining columns rolling averages of features of all articles published within some time period, multiplied by the length of the offer term, are also included. "Intro price" is the price, in dollars, of the lowest-priced offer in the menu shown to users in the cell. Prices are in absolute terms and not scaled by offer length.

As soon as we add the scaled-by-offer-length rolling-average features, we again recover a negative price sensitivity as expected. This conclusion holds regardless of the window length chosen, but because the price sensitivity estimate is largest (in magnitude) and most precise at a window length of two weeks, we use that length in the demand model.

B.1.4 Matching on Reading History

Both our reduced form regressions and structural model of demand rely on the identifying assumption that conditional on our discretized measure of the user's intensity of visits to the site, the specific article on which the user encounters a paywall is random. This assumption would be violated if there were segregation among users in the kinds of articles they read, within our discrete user type bins. For example, suppose there were some heavy readers who always read sports articles and some equally heavy readers who always read politics articles, and imagine that the sports readers have systematically higher propensity to subscribe than do politics readers. Our method would group these types together, and assign higher value to sports than to politics articles, not because these articles are actually more valuable to all readers but because they are more likely to be read by a group of readers with relatively high subscription propensity.

As a test of our identifying assumption, we split our high-readership (bin 3) types into "hard news" and "soft news" readers, by computing the average feature vector of the previous three articles read

[40] This is how users in our demand model form forecasts about the future articles they expect to be able to read while subscribed.

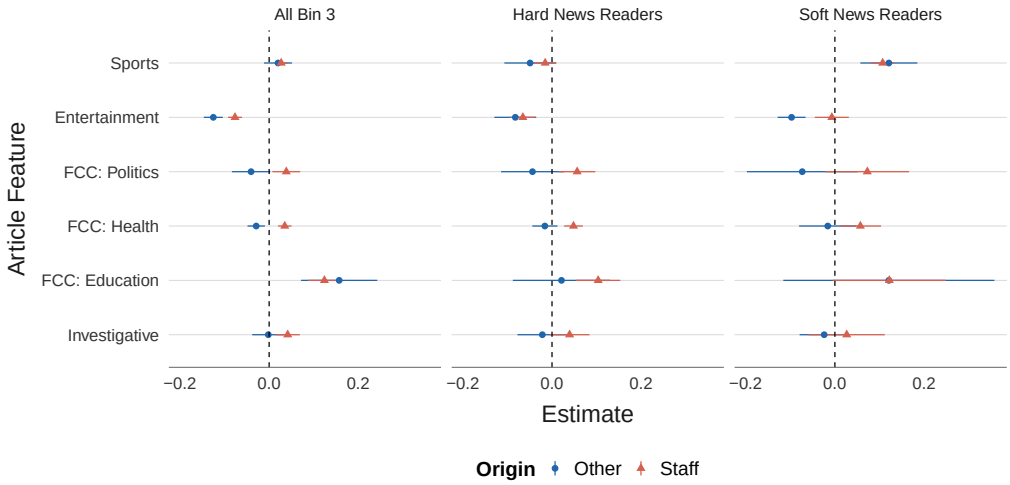


Figure B.4: Split-sample subscription regressions on readers who previously read hard news vs. soft news articles.

by each bin-3 user at each point they encountered a paywall. "Hard news" readers are those whose last-three-article average feature value of Civics, Economy, Education, Public Safety, Environment, Health, Politics, or Investigative exceeded 0.05, and whose last-three-article average feature value of Columns, Entertainment, To Do, and Sports were all less than 0.05. "Soft news" readers are defined using the obverse criterion.[41] These groups are approximately equally sized, with just under 4M user-paywall observations in each.

We then computed our reduced form subscription regressions on only the hard-news or soft-news readers alone. Figure B.4 shows the results of this exercise for a subset of the article features, along with the corresponding results in the full sample of bin 3 readers for comparison. Although the estimates on the subgroups are noisier, the overall pattern is quite similar for both groups.

B.2 Demand model estimates

This section provides estimates of parameters of the demand model described in Section 3. First, Figure B.5 displays estimates of β_x^V for each user type. These parameters govern the number of users of each type predicted to visit an article, as a function of that article's features. There is a clear pattern that articles written by in-house staff generate more traffic from all user types than do outsourced articles, consistent with the reduced form.

Second, Figure B.6 displays estimates of β_x^R for each user type. Again, in-house articles generally outperform outsourced articles. For users of type 2 and 3, the CIN categories generally have more positive impact on users' subscription propensities than non-CIN categories. Users of type 1 almost never subscribe, regardless of article content.

[41] Note that the groups are not mutually exclusive: there are readers who fall into neither group.

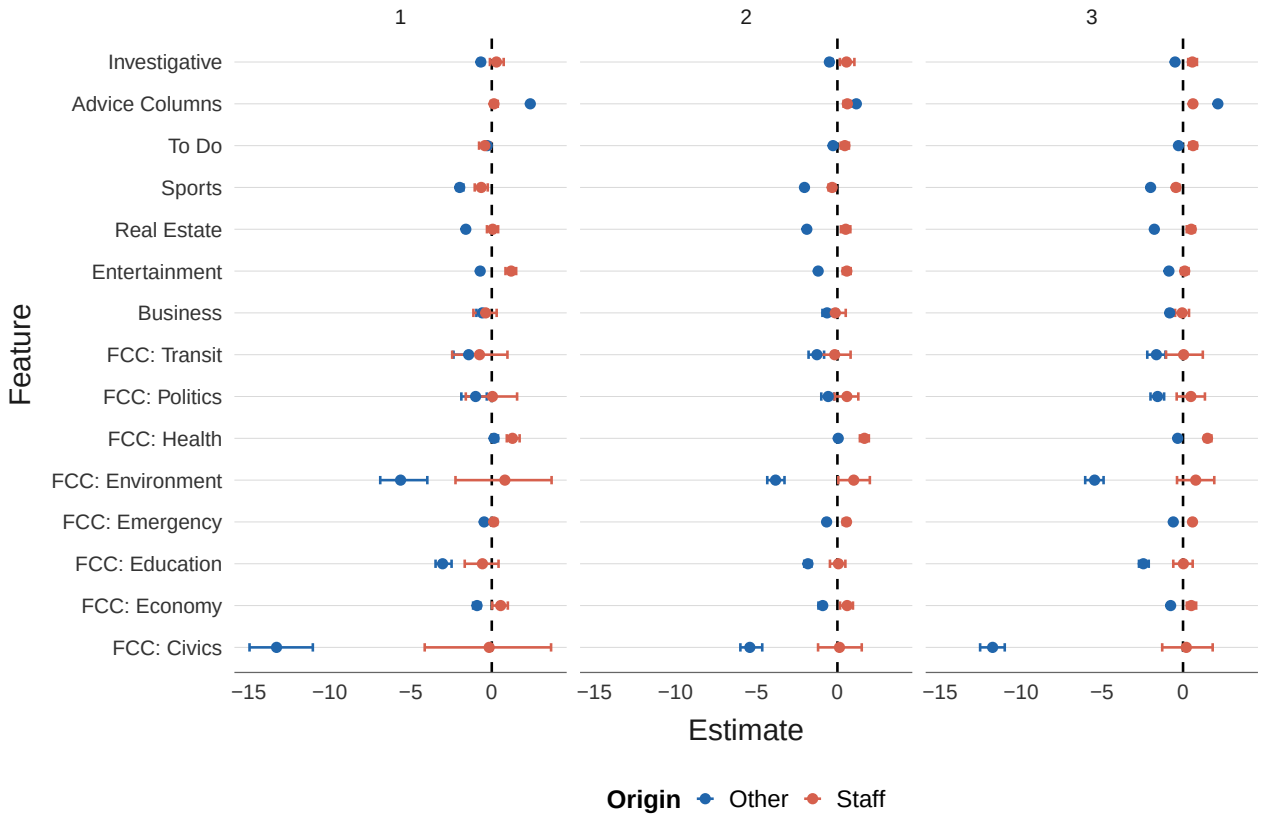


Figure B.5: Effects of article features on user arrivals. Points are the maximum likelihood estimate of β_x^V, the effect of the feature on the article's Poisson mean parameter governing arrivals from each user type $x \in \{1, 2, 3\}$. Confidence intervals are the central 95% interval generated by 100 bootstrap replications, according to the method described in Section 3.1.

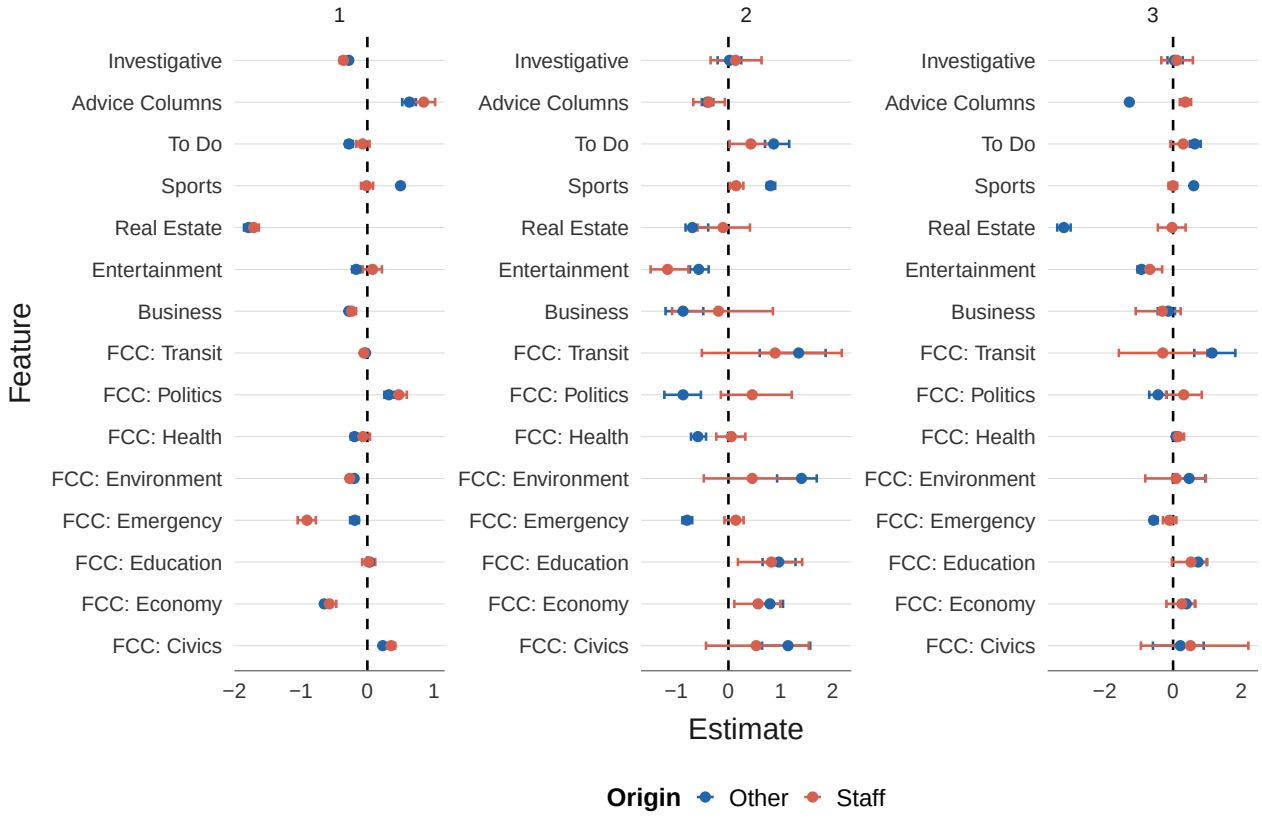


Figure B.6: Effects of article features on subscription probabilities, conditional on a user visiting the article. Points are the maximum likelihood estimate of β_x^R, the effect of the feature on the article's value to the reader for each user type $x \in \{1, 2, 3\}$. Confidence intervals are the central 95% interval generated by 100 bootstrap replications, according to the method described in Section 3.1.

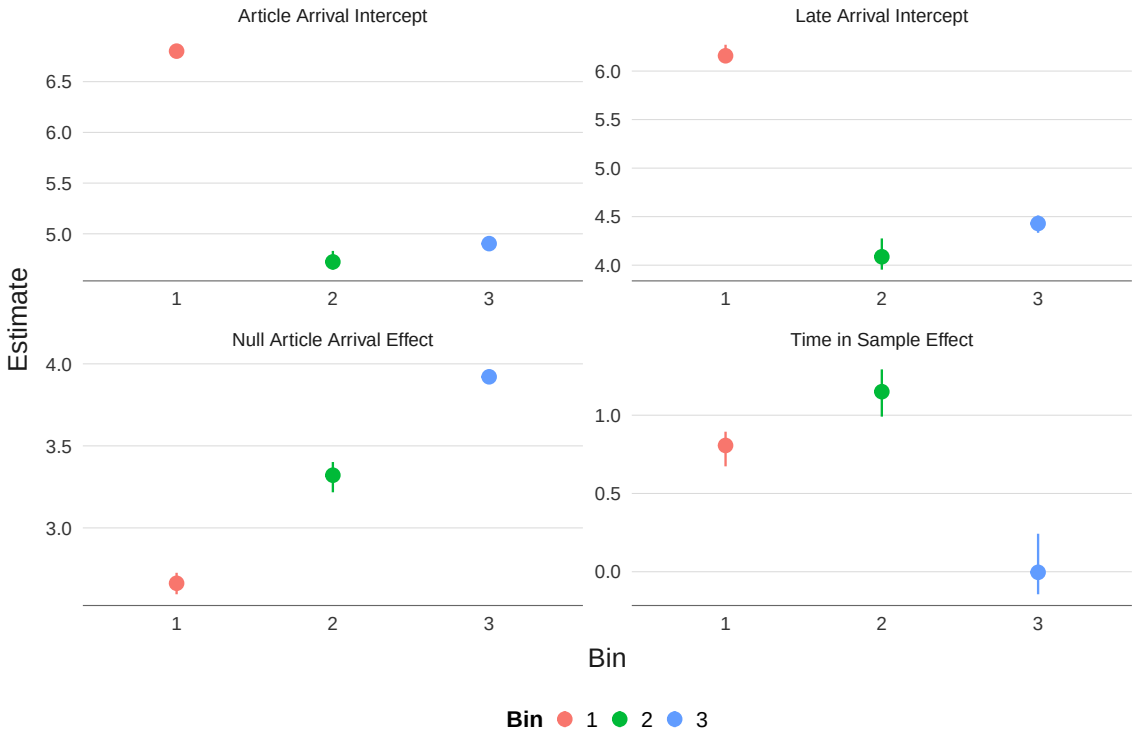


Figure B.7: Additional arrival model parameters. Points are the maximum likelihood estimates of $\xi_x^V, \xi_x^{LV}, \xi_x^{NV}, \tau^x$ for each user type $x \in \{1, 2, 3\}$. Confidence intervals are the central 95% interval generated by 100 bootstrap replications, according to the method described in Section 3.1.

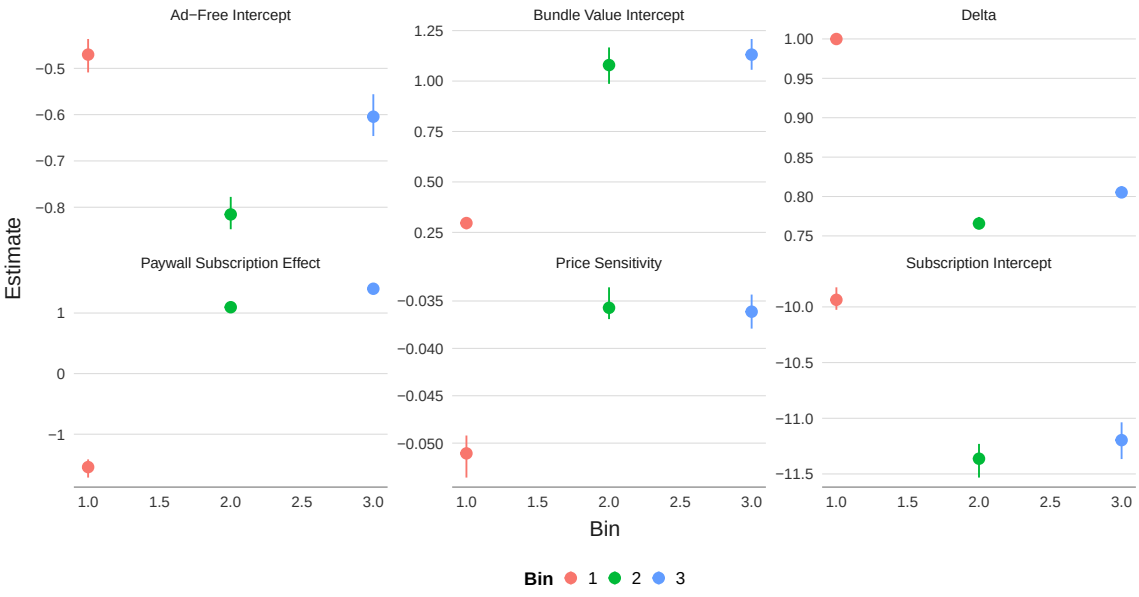


Figure B.8: Additional subscription model parameters. Points are the maximum likelihood estimates of $\kappa_x, b_x, \delta_x, \eta_x^R, \alpha_x$, and ξ_x^R, for each user type $x \in \{1, 2, 3\}$. Confidence intervals are the central 95% interval generated by 100 bootstrap replications, according to the method described in Section 3.1.

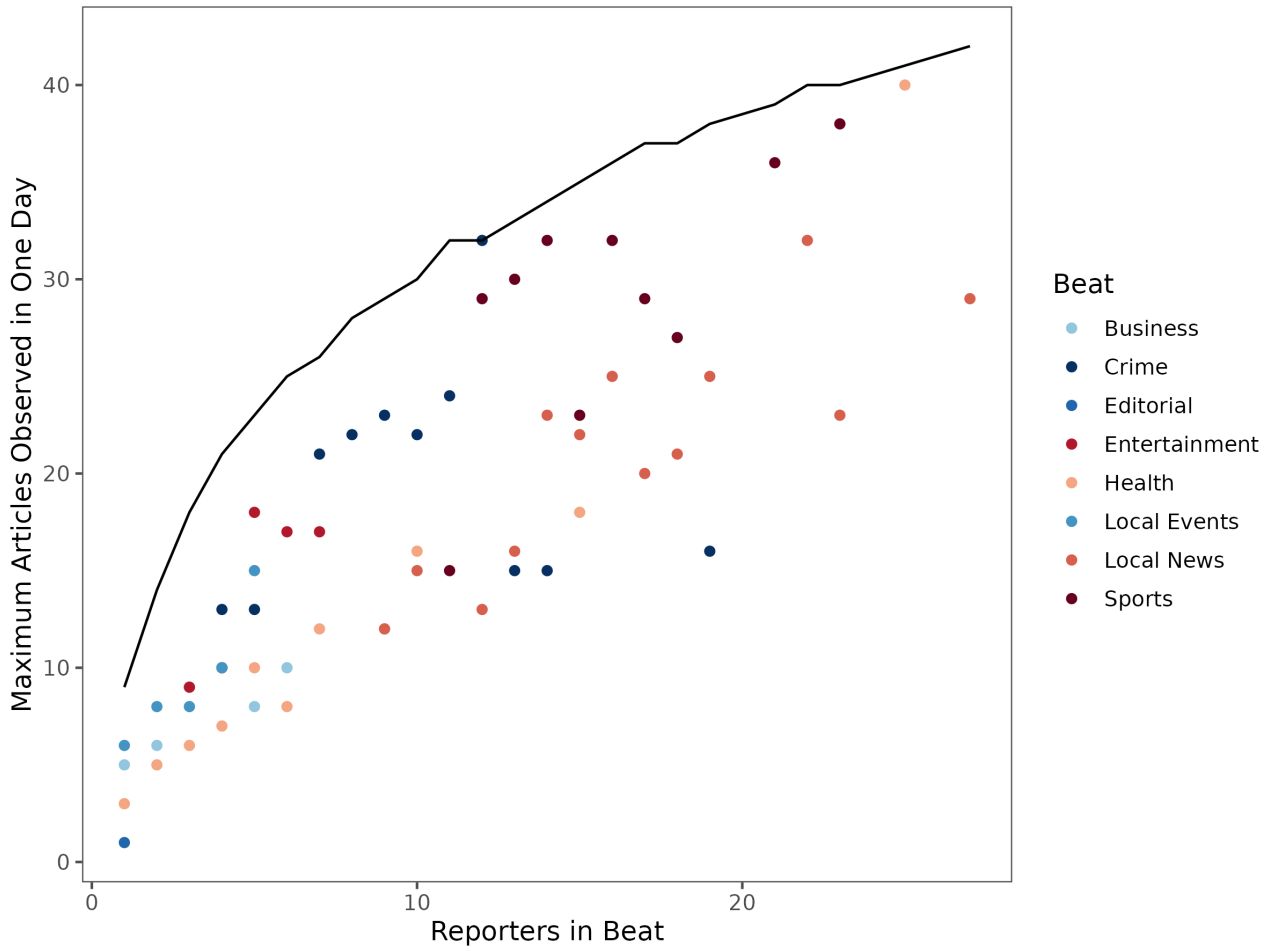
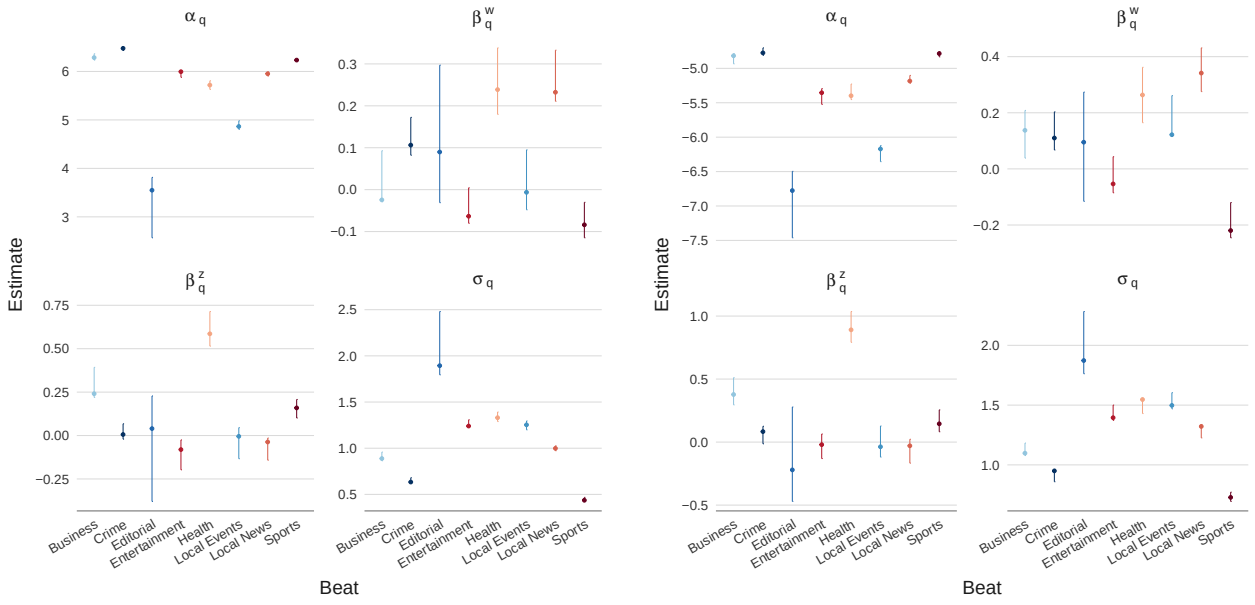


Figure B.9: Estimated relationship between number of reporters and maximum daily articles produced. Points are the maximum number of articles ever observed to be produced in one day for a given beat with a given staffing level. The line is the fitted curve, ceiling$(\gamma^{\bar{N}} \log(1 + A_{b,d}))$*, with* $\gamma^{\bar{N}} \approx 12.71$.

B.3 Supply model estimates

Figure B.9 visualizes the calibration of $\bar{N}_{b,d}$ using data on the observed maximum daily production of articles at different staff levels. Figures B.10 and B.11 present estimates with bootstrapped 95% confidence intervals for the parameters of the supply model detailed in Section 4.



(a) Visits.

(b) Subscriptions.

Figure B.10: Estimates and bootstrapped standard errors of supply model parameters that control the quality distribution of published articles. The left panel is visit quality $\hat{q}_a^V$ and the right panel is subscription quality q_a^R. α_q is the constant term, and β_q^w and β_q^z are multipliers on the beat-specific and overall-newspaper Google trend scores respectively. σ_q is the standard deviation of the quality distribution.

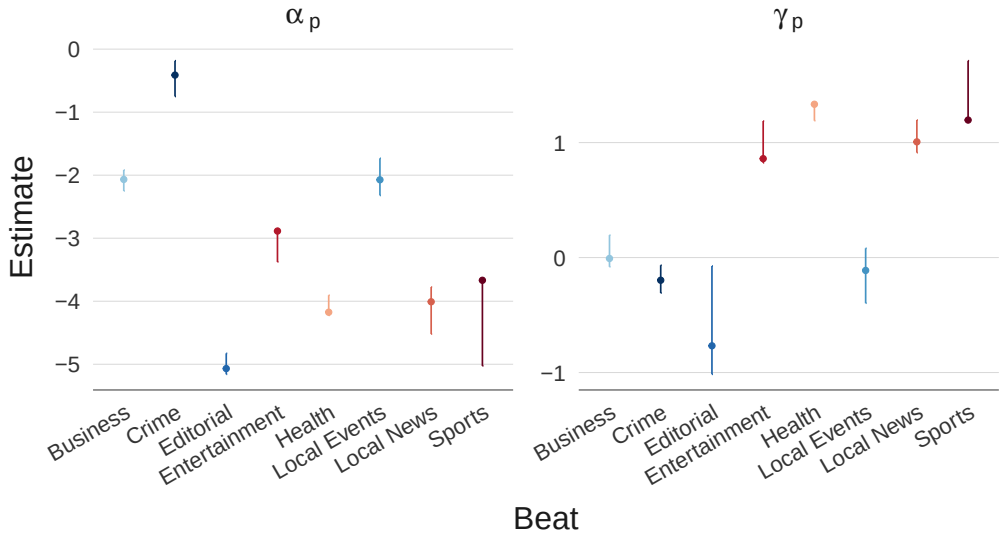


Figure B.11: Estimates and bootstrapped standard errors of supply model parameters that control the quantity of articles published as a function of staffing. α_p is the constant term and γ_p is the multiplier on log staff in the production equation.