What do news readers want?*

Gregory J. Martin† Cameron Pfiffer‡
Shoshana Vasserman§

PRELIMINARY AND INCOMPLETE – PLEASE DO NOT DISTRIBUTE OR CITE

November 2024

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 from 2020 to 2023, we investigate consumers' willingness to pay for different kinds of content, including that believed to have the greatest civic externalities. Our identification 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 content portfolio for the paper under different counterfactual revenue models: a fully subscription-based model and a fully ad-supported model. Preliminary results suggest that news consumers are willing to pay for investigative and local content, and that measures of demand based only on time-use substantially underestimate the value of investigative news.

*We are grateful to an anonymous local news organization for making this project possible.
†Stanford GSB. Email: gjmartin@stanford.edu.
‡Stanford GSB. Email: cpfiffer@stanford.edu
§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 the 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., 2024; 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, both at 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. 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, this 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, deeply researched reporting of the kind taught in journalism schools or rewarded in journalism prize competitions — within media organizations led to the production of more of what is variously called investigative or accountability journalism than readers truly wanted (Hamilton, 2004).

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. Combined with the rise of new measurement technologies that give management at media organizations more visibility into the correlates of reader attention, many observers fear a permanent shift of news production in the direction of shorter, "clickbait" articles 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 might consume so-called "hard news" if forced to do so by the limitations of the distribution technology, but will opt out 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 relying on real-time metrics (Christin, 2020) necessarily lead to the elimination of investigative reporting, as managers realize the weakness of consumer demand for such pieces?

Our method relies on very fine-grained 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 media market resulting from the expansion of broadband internet access, legacy metropolitan dailies like this one still produce a disproportionate share of coverage satisfying public information needs, particularly at the local level (Mahone et al., 2019; Hamilton, 2016).[1] We are able to track reader decisions to read and to subscribe, or not, at the individual-by-article level, and can compare reader willingness to pay for such socially-valuable content compared to articles (sports coverage, for example) with more purely 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 available rich article-level characteristics based on article 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

[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)."

subscription rate between paywalled and non-paywalled users identifies reader willingness to pay for article attributes. Once we account for differences in users' baseline propensity to be paywalled (which correlates with frequency of visits to the site) residual variation derives from small differences in the timing of visits or the specific article selected at the time of crossing the paywall threshold.

The resulting demand estimates can inform two distinct areas of policy uncertainty. First, for news outlets, there is a live question of how best to reorganize the business model in the wake of the disruption experienced since the turn of the century. It is clear enough that the old model 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 revenue implications of several alternative models, including a pure advertising-based model, a monthly subscription model, and an a la carte model. We additionally examine how the optimal content bundle would change under these variants. We find 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 at opportunities to unlock access to articles related to areas of critical news coverage like public health, economics and local politics.

For regulators and the public at large, the provision of these special kinds of news coverage have 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 required.

1 Background and Data

Our dataset comes from a metropolitan daily newspaper headquartered in a large US city.[2] The newspaper is characteristic of many legacy papers in the US in that it enjoyed a local monopoly during the print era, has experienced a large decline in print subscription and display advertising revenue,[3] and has refocused on growing its digital subscription and advertising businesses. It also, like many US papers, is currently owned by a private-equity-controlled holding company.[4] The paper's audience is mostly but not exclusively regional — according to the Alliance for Audited Media, a slight majority of its online readers reside in the media market (Designated Market Area or DMA) in which

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

[3] 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.

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

the paper is headquartered.[5]

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.[6] 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 term 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 his / her engagement with the article (e.g. how far he or she 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.[7] We observe a total of about 1.2B user-sessions in our data and 605M article visits.[8]

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.[9] 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[10] 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, he or she is redirected to the home page and is unable to access the article. If he or she subscribes, she can continue to read the article, in addition to any other articles he or she wants to read in (at least) the next month.

[5] 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.

[6] 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.

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

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

[9] 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.

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

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.[11]

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 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.

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

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

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). 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 not identified by the FCC as critical needs. 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[12] 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 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.[13] Bin 3 users are more than 100 times as likely to subscribe as those in bin 1, conditional on encountering a paywall.

[12] 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.

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

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

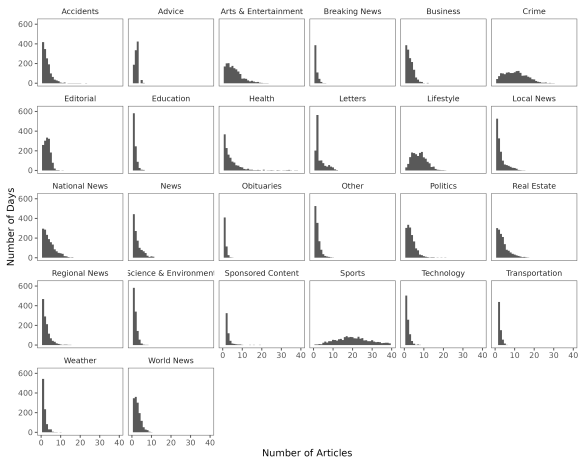
We first present several stylized facts that characterize the important dimensions of variation that we observe in the data.

2.1 Production

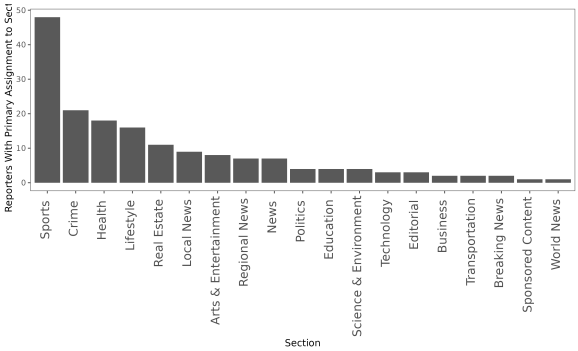
The newspaper offers a mixture of content across a variety of topical categories. Figure 1a shows histograms of the number of articles published on a single day from a given section. Some sections, like Sports, have high mean but also high variance, as coverage intensity tracks events like playoffs or 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.

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



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)[14] across all

[14]The supervised-learning-based features are continuous, but their distributions are all highly bimodal with peaks near

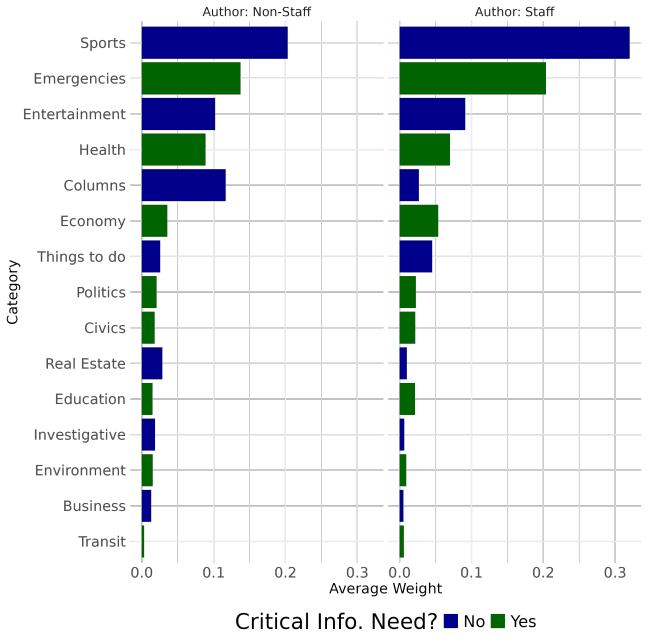


Figure 2: The average weight across all articles in the dataset, split by articles written by in-house staff (right) and by other authors such as wire services (left). Categories identified by the FCC as "Critical Information Needs" are highlighted in green.

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. A.1 in the Appendix shows the correlation matrix between each of our features.

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.[15]

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

0 and near 1.

[15] 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.

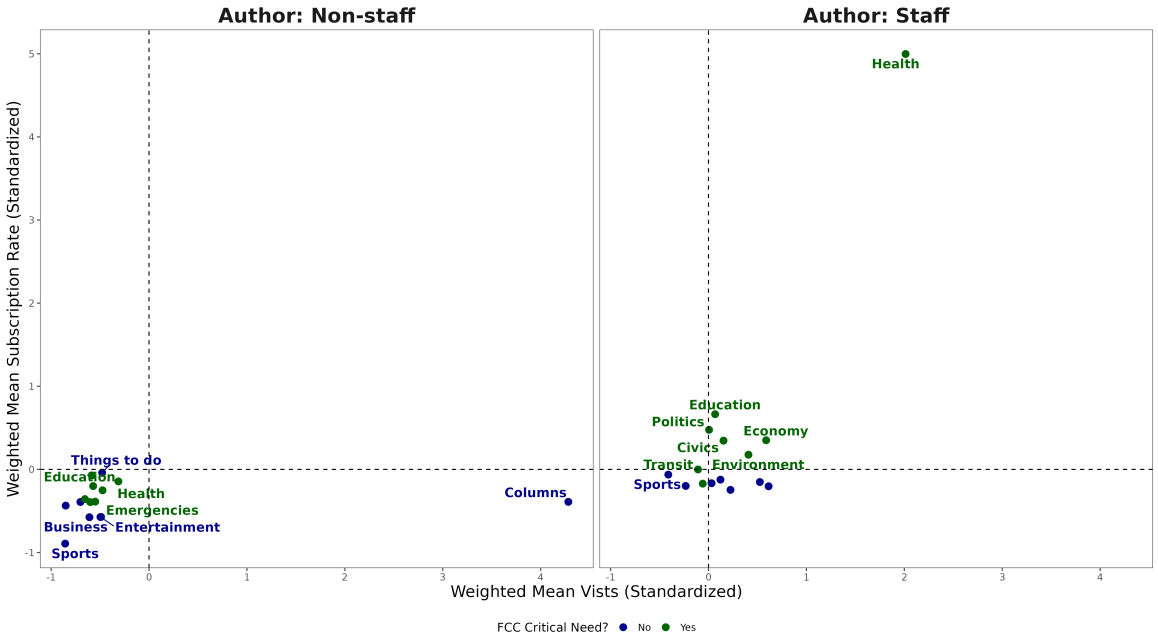


Figure 3: Visits versus subscriptions, by feature. The y-axis is the average value of the subscription rate for users paywalled on each article in the dataset, weighted by that article's value of the indicated feature. The x-axis is the average value of the number of unique visitors for each article 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.

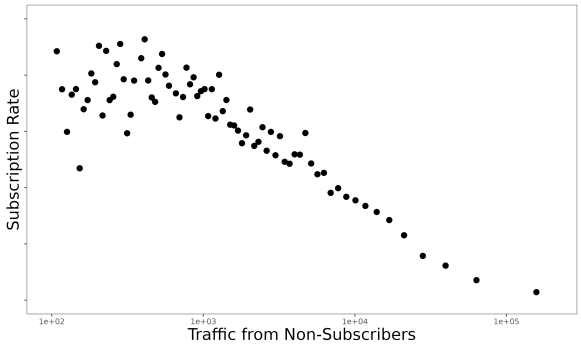
On the subscription side, in-house articles also almost uniformly produce higher subscription rates than do outsourced articles. Within in-house articles, there is substantial variation in subscription rates. The pattern is generally that categories with higher social value (those identified by the FCC 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 A.3 show that these conclusions generally hold in a regression setting as well.

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 A.6) could be biased if some articles bring in visitors who have different baseline subscription propensities.[16]

[16]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.

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

(a) Subscription rate.



(b) Traffic from lowest PC_1 bin.

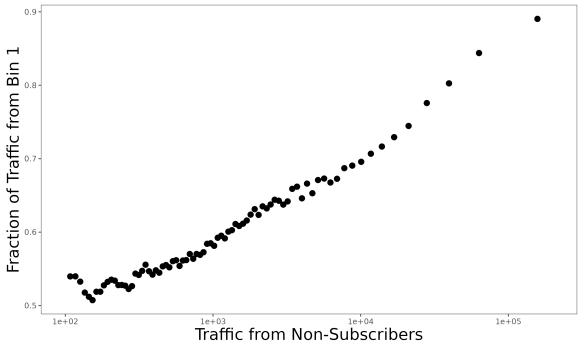


Figure 4a shows that there is indeed a correlation between total traffic to an article and the fraction of paywalled readers who subscribe. Each point in the figure is the average of all articles within the same percentile of the logged traffic distribution. The figure shows that the correlation is fairly 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. The analysis thus far ignores prices, but as described previously, there is substantial variation in the terms offered to users over our sample period. We investigate here how sensitive users are to the offered prices and subscription terms. As a simplification that collapses the two primary dimensions of offer variation (intro price and intro length) into one, we in this section consider 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. We match each paywall event to the menu code that the user saw, and use the lowest average price offered among all options in that menu as the right-hand-side variable in the regression.

Table 2 shows that users are responsive to price changes, and subscribe less when offer terms are less favorable. The highest-engagement users (bin 3) are the most price sensitive, though this is 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	$-2.64 \times 10^{-5***}$	$-3.23 \times 10^{-7*}$	$-1.52 \times 10^{-5***}$	$-4.14 \times 10^{-5***}$
	(2.43×10^{-6})	(1.66×10^{-7})	(2.78×10^{-6})	(4.66×10^{-6})
Visit Date FEs	Y	Y	Y	Y
Observations	13,495,145	4,874,967	5,751,063	2,869,115
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

Finally, 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 are 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), in 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 subscription 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 significantly. The point estimate for the investigative indicator is also negative, although the base level is so low that the change is not significant here. Per Table 3, these declines occur as the user reads more articles in total, implying dilution of these kinds of articles by other content.

These changes, consistent with Figure A.6, indicate that these features are *positively* associated with users' valuation; when unsubscribed and reading opportunities are scarce, users allocate a greater fraction of their reading time to this content than they do when reading is unconstrained. Conversely, the average level of less-socially-important features like Columns and Entertainment goes up after subscription, indicating that these kinds of articles have lower value.

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	185,997	185,997	185,997	185,997
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. N's are omitted per our data sharing agreement which prohibits releasing subscriber numbers. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

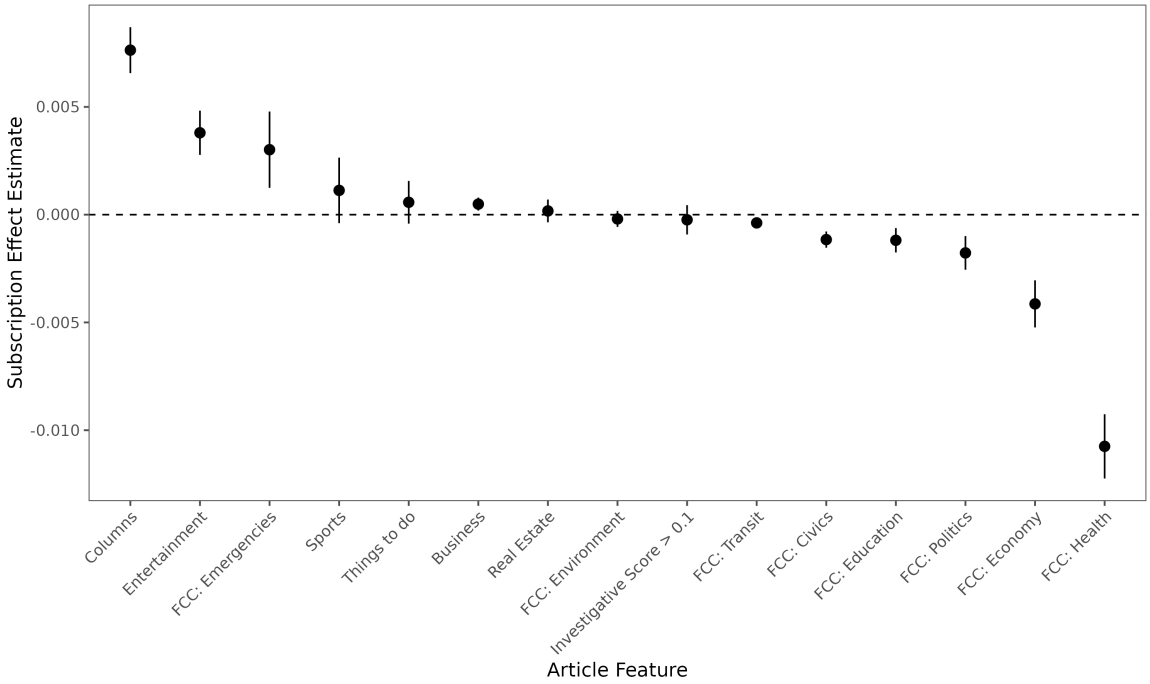
3 Demand Model

In this section, we present a model of consumer 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 subscriptions offers; and (4) readers may be more likely to subscribe when they are paywalled from certain types of articles.

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

Articles Each day d, the newspaper publishes a new portfolio 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 features described in Section 1.

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.



Consumer Visits to Articles Consumers 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 potential consumers arrive at an article a on day d according to a Poisson distribution parameterized as follows:

$$V_{a,d,x} \sim \text{Poisson}\left(\exp\left(\xi^A_{x,d} + \beta^A_x \theta_a\right)\right).$$

Here, $V_{a,d,x}$ is the number of visits to article a on day d by a readers of type x. The Poisson rate for $V_{a,d,x}$ is given by a generalized regression function of article characteristics θ_a with a reader-type $\times$ day-level intercept (ξ^a_d) under an exponential link. Note that our model assumes that any individual reader could visit any collection of articles. This reflects the observation that readers often peruse multiple articles once on the newspaper's website. However, we do not model explicit correlations between article visits.

The vast majority of article visits correspond to recently published articles. For articles published more than three days prior to an observed visit, we model the total number of late arrivals with a modified Poisson regression, replacing the reader-type *times* day fixed effect ξ^a_d with a reader-type-specific late-arrival fixed effect ξ^{LA}_x and a linearly additive control for the number of days $\tilde{D}_a$ that article a was available for reading during our panel window:

$$V^{LA}_{a,x} \sim \text{Poisson}\left(\exp\left(\xi^{LA}_x + \gamma^{LA}_x \tilde{D}_a + \beta^A_x \theta_a\right)\right).$$

Paywalls Consumers 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 browser error. We assume that paywall events are random conditional on the users' observable characteristics at the time of their visits, and are independent of the attributes of the articles being visited.

Consumer Values for Articles Consumers only benefit from an article a if they *both* visit the article page *and* are able to read the article. That is, consumers receive no value from visiting an article if they hit a paywall and are blocked from accessing the article's contents. For simplicity, we assume that a consumer of type x receives a utility of $\beta_x^R \theta_a$ if she visits article a and is not blocked from reading it due to a paywall. If the consumer does hit a paywall, then she receives a utility of 0 unless she purchases a subscription offer.

Subscription Offers On each visit, consumers receive a menu of promotional subscription offers. Each offer o consists of three components: a duration T (in months) during which the consumer would have unrestricted access to the newspapers' online content, a per-month price p, and a quality q.[17] To evaluate an offer o, consumers consider both the immediate value of subscribing, and the net present value of being subscribed over the duration of the offer. If the consumer is blocked from reading an article a by a paywall, then the immediate value of subscribing is her value for the article $\beta_x^R \theta_a$. If the consumer 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 consumers are forward looking and can accurately predict the articles that will be published over the duration of each offer. This assumption reflects the idea that many types of reporting are predictable—either because they cover scheduled events such as elections and sporting events, or because they involve ongoing coverage of an evolving story. To goal of the bundle value in our model is to both capture the effect of this predictable, time varying news coverage, and account for the difference in utility that subscription offers with different durations entail. To this effect, the net present value of an offer o considered on day d by a reader of type x is given by:

$$B(o, d, x) = \sum_{t=1}^{T_o} \sum_{a \in A_{d+t}} \delta^t \beta_x^R \theta_a, \tag{1}$$

where δ is a daily discount factor and A_{d+t} is the set of articles published t days from the date on which the offer is evaluated.

In addition, offers can be "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 $\kappa_{o,x}^R$.

[17] For now, we omit the renewal price and assume that consumers do not renew their promotional subscriptions.

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 the ad-free offer alone. In many cases, we can observe the specific menu that was shown to users along with a series of 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 its observed empirical frequency.

Subscription Propensity Consumers faced with a menu of subscription offers can choose to purchase one of the offers, or not to subscribe. We assume that consumers 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, offer-reader-type and day-level intercept, 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^R. Finally, we assume that consumer utility is quasi-linear in prices, and subject to a price coefficient α_x.

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

$$u_{i,a,d,x}(o) = \xi_{x,d}^R + \kappa_{o,x}^R - \alpha_x p_o + \mathbf{1}\{\text{paywall}\} \cdot \left(\eta_x^R + \beta_x^R \theta_a\right) + B(o,d,x) + \varepsilon_{i,o,d,x}. \tag{2}$$

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

$$\pi_{i,a,d,x}(o) = \frac{\exp(\xi_{x,d}^R + \kappa_{o,x}^R - \alpha_x p_o + \mathbf{1}\{\text{paywall}\} \cdot \left(\eta_x^R + \beta_x^R \theta_a\right) + B(o,d,x))}{1 + \sum_{o' \in O_i} \exp(\xi_d^R + \kappa_{o',x}^R - \alpha_x p_o' + \mathbf{1}\{\text{paywall}\} \cdot (\eta_x^R + \beta_x^R \theta_a) + B(o',d,x))}. \tag{3}$$

4 Estimation

Our model produces three sets of parameters to estimate: (i) the article visit day effects $\xi_{x,d}^A$, article characteristic coefficients β_x^A and late arrival parameters ξ_x^{LA} and γ_x^{LA}; the subscription day effects $\xi_{x,d}^R$, article characteristic coefficients β_x^R, the price disutility coefficients α_x, and fixed effects $\kappa_{o,x}^R$ and η_x^R; and (iii) the daily discount factor δ. To estimate these parameters, we maximize the joint likelihood of site arrival, article arrival and subscription purchasing probability for each visit in our data.

5 Results

Our primary interest is in the relationship between newspaper revenue and editorial choices about content priorities. Understanding the effects of content on revenue is complicated by the fact that our demand model allows article content to affect revenue through three channels: the amount of traffic that the article generates, the value of the article to readers paywalled on that article, and the contribution of the article to the overall bundle value of all content available to subscribers over a given subscription term. An article feature with low reader value (β^R) may nonetheless be an important driver of subscription revenue if it generates a lot of traffic (i.e., has high β^A).

To synthesize the effects across all three of these channels, we present results in Figure 6 of an exercise in which we use the fitted model to predict the change in total traffic to the newspaper and in subsequent subscriptions over our sample period that would result if an article feature were entirely eliminated from the newspaper's production. That is, for each column k in the article feature matrix Θ, we set Θ_k to zero; predict traffic generated by all articles after this modification; and finally predict subscriptions resulting from these visits given the modified article characteristics. This calculation holds all other features and all prices and subscription offer characteristics constant. In computing changes in subscriptions, this exercise combines the direct effect of article features on reader value (β^R) with the indirect effect through the number of readers arriving to the site and facing the option to subscribe or not (β^A). Full estimates of both parameter vectors are provided in Appendix A.4, Figures A.7 and A.8.

The results show that outsourced articles contribute little to traffic to the paper. The sole exception is syndicated advice columns, without which the newspaper's total traffic would fall by about half a percentage point over this period. The effect on subscriptions of eliminating article features in non-staff-written articles are generally positive and sometimes substantially so; this reflects the fact that we estimate the value of these kinds of articles to be generally negative, and when eliminated the overall value of the newspaper's bundle (B, in the demand model notation) rises.

Eliminating staff-written article features, on the other hand, nearly always leads to declines in both traffic and subscribers, some of which are substantial. There is a general pattern that eliminating CIN features (labeled in green in the plot) produce much larger declines, particularly in subscriptions, than eliminating the non-CIN features. The exception is Sports, which makes up a large enough share of articles and is responsible for generating enough traffic to make up for the fact that its reader value coefficient is small relative to the more socially-valuable categories.[18]

[18] See Figure A.8; the value of β^R for all reader types is lower than that for Politics, Civics, or Economics.

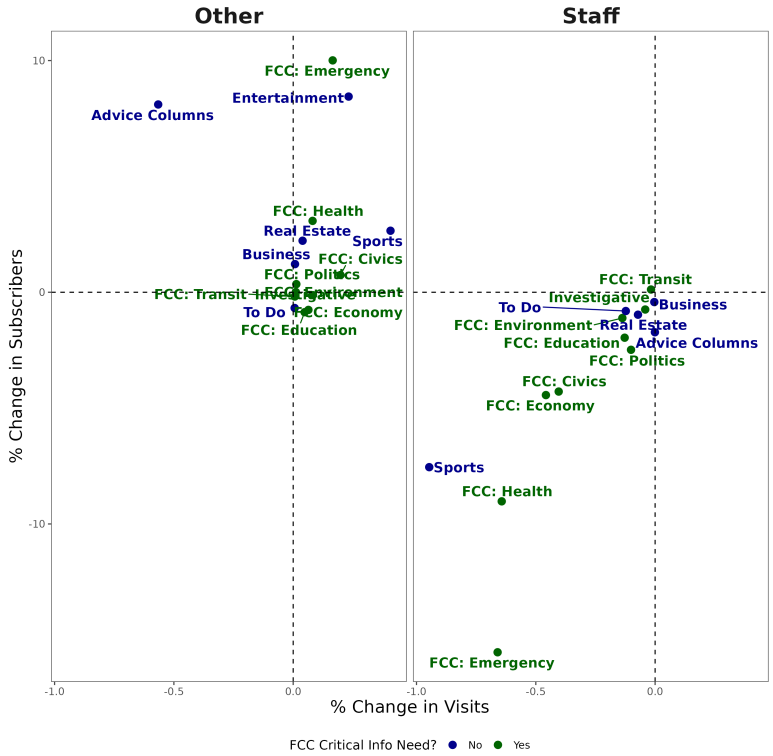


Figure 6: Estimated effects of eliminating a feature (setting the value of that feature for all articles to 0) on total traffic (horizontal axis) and total subscriptions (vertical axis). All other article features, and all prices, menus and subscription offers, are held constant. We plot changes relative to the predicted values from the demand model at baseline. Negative values indicate that traffic or subscriptions fall relative to baseline when the feature is eliminated.

6 Conclusion

Coming soon!

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 (2024): "The Impact of Online Competition on Local Newspapers: Evidence from the Introduction of Craigslist," *Review of Economic Studies*.

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. (2004): *All the news that's fit to sell: How the market transforms information into news*, Princeton University Press.

Hamilton, J. T. (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;[19] 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

[19] 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.

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.[20] 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 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 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.

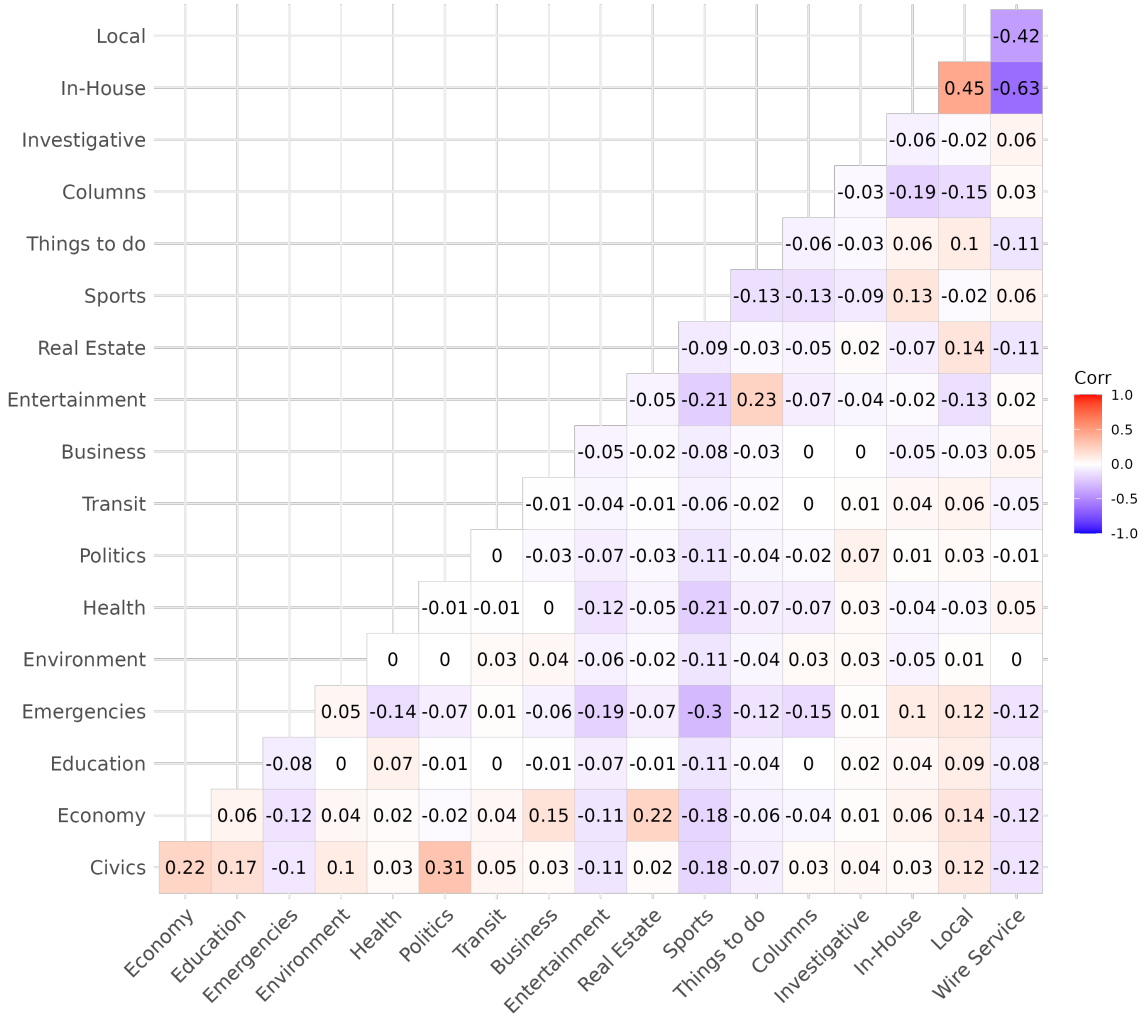
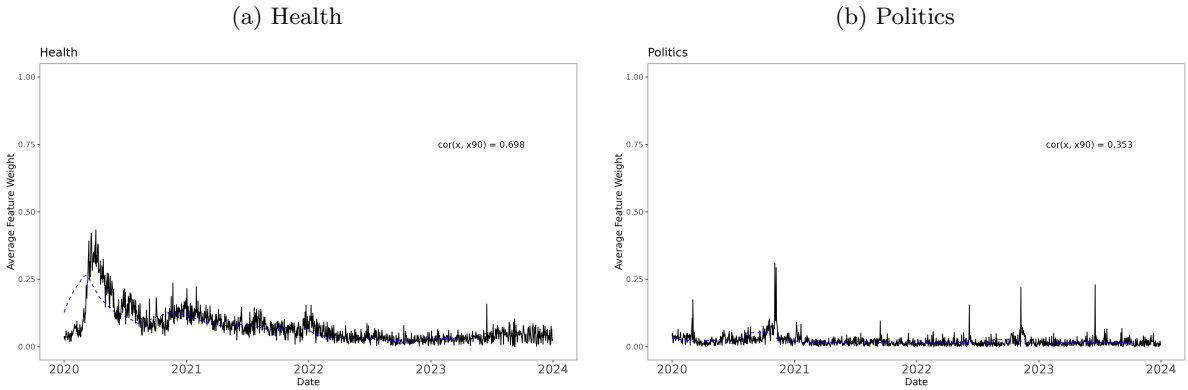


Figure A.1: Correlations between Article Features.

Figure A.2: Time trends of the average feature weight in the Health and Politics categories. The black lines are the daily average, and the blue dashed lines are the forward 90-day rolling average. The reported correlations are the correlations between the daily and forward 90-day rolling averages.



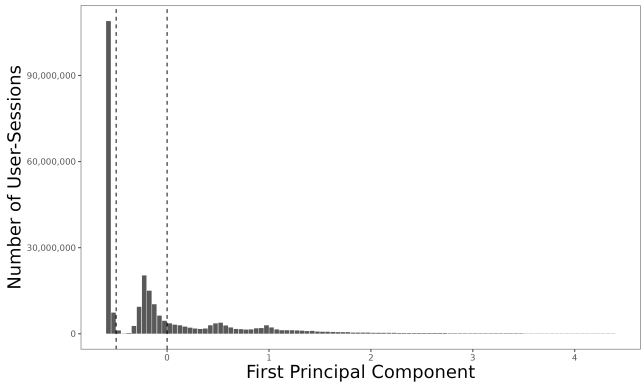


Figure A.3: Histogram of First-Dimension Principal Component scores (PC1). This the score is derived from principal components analysis applied to a matrix consisting of a set of variables measuring the user's intensity of interaction with the site in the past six weeks, described in Section 1. PC1 varies 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, histogram is truncated at 5; there are an additional 7.76M user-sessions with PC1 greater than 5 not plotted.

A.3 Additional reduced form traffic and subscription regressions

Figure A.5 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 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

[20] 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.

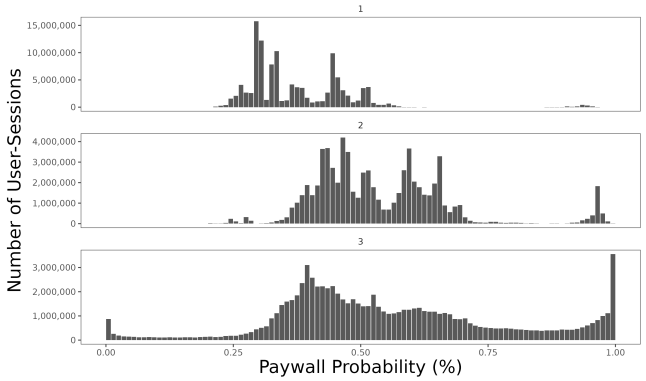


Figure A.4: Histogram of estimated propensities of encountering a paywall by user-article visit. The histogram is split by user type (1,2, or 3) defined by the user's PC1 score as described above.

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.

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 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 A.6 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.[21] 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 A.5 shows is the 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

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

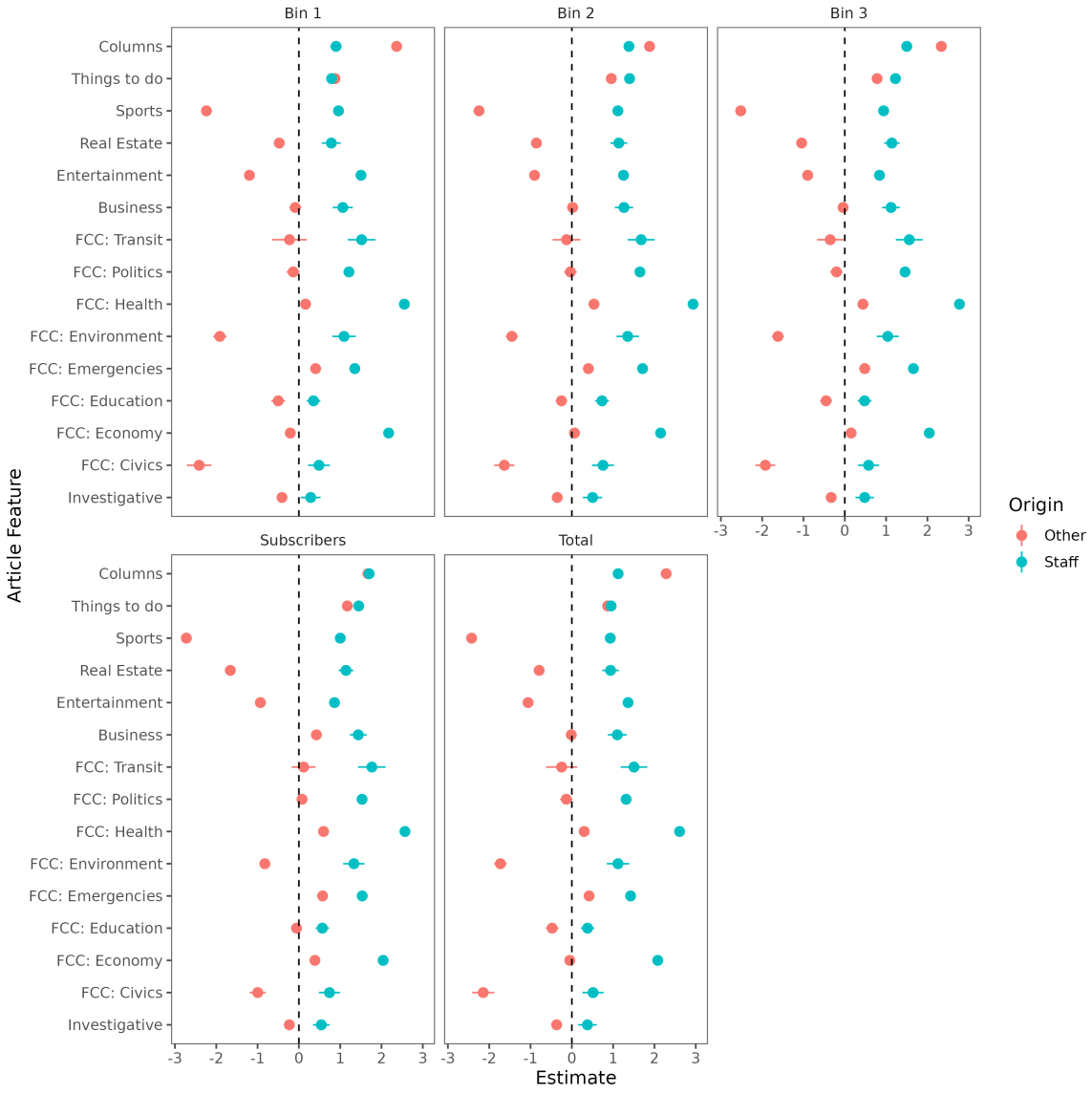


Figure A.5: Correlates of article traffic. Each panel shows the coefficient and its 95% confidence interval of the indicated feature in a model of IHS-transformed visits to the article within 14 days of publication. In the top three panels, the outcome is visits by each non-subscriber subtype; the bottom 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.

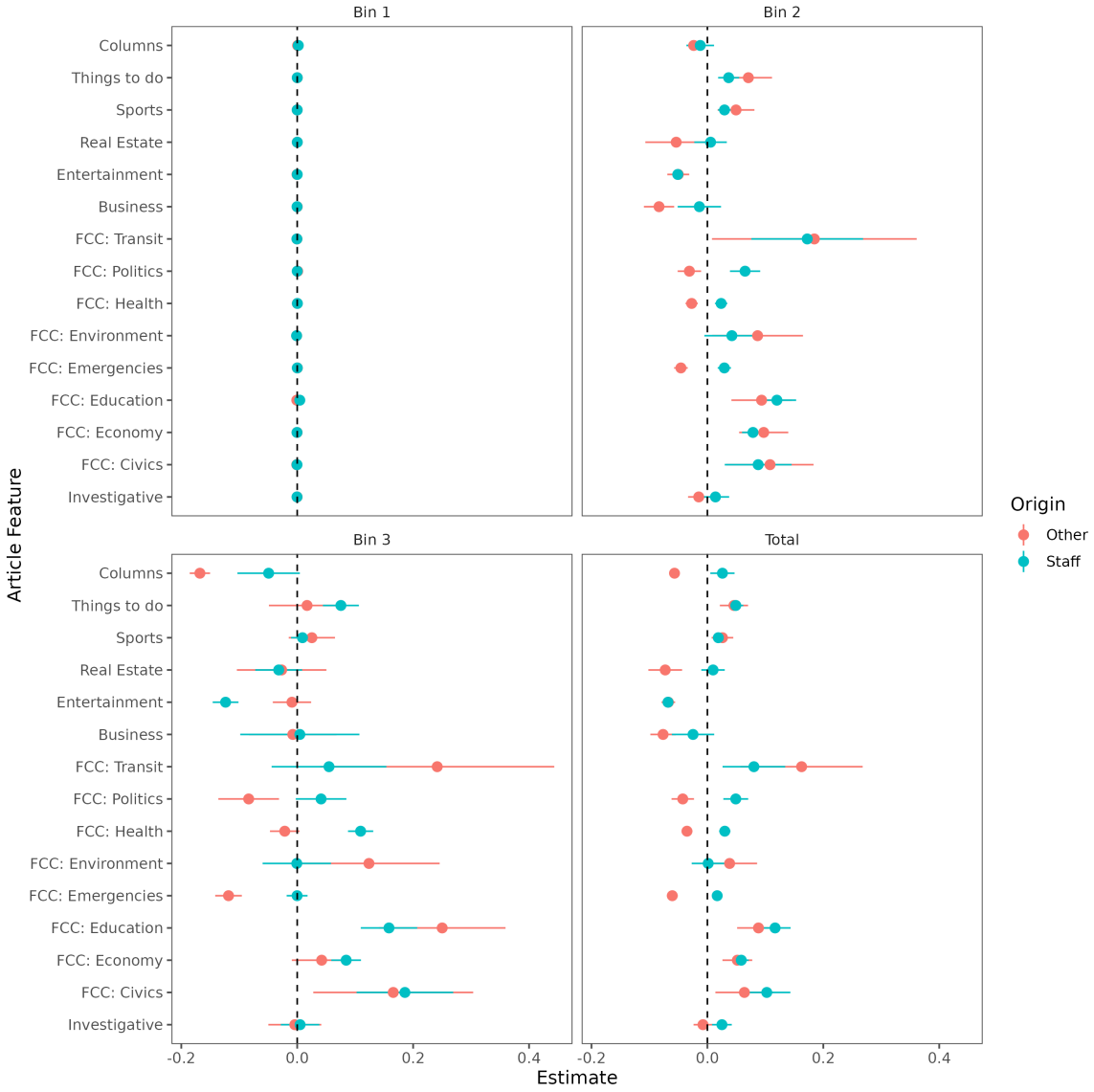


Figure A.6: Correlates of subscription, conditional on the features of the article on which a user encountered a paywall. Each panel shows the coefficient and its 95% confidence interval of the indicated feature in a linear model of subscription rate, conditional on hitting a paywall. An observation 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.

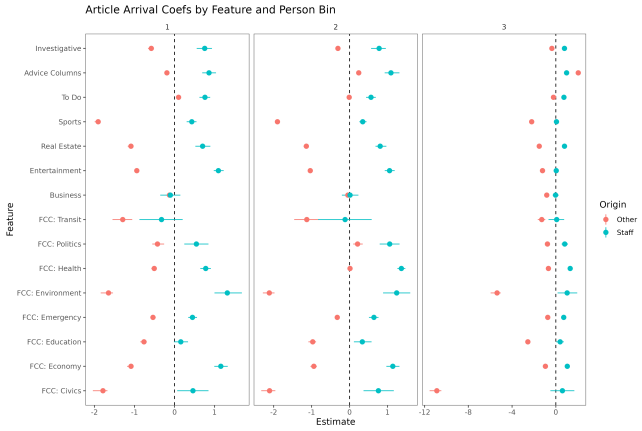


Figure A.7: Effects of article features on user arrivals. Points are the maximum likelihood estimate of β_x^A, the effect of the feature on the article's Poisson mean parameter governing arrivals from each user type $x \in \{1, 2, 3\}$. Confidence intervals are the central 95% interval generated by 100 bootstrap replications, according to the method described in Section 4.

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

A.4 Demand model estimates

This section provides estimates of parameters of the demand model described in Section 3. First, Figure A.7 displays estimates of β_x^A 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 A.8 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.

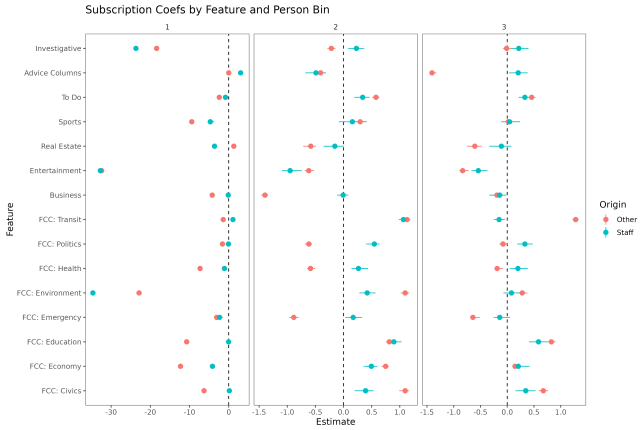


Figure A.8: Effects of article features on subscription probabilities, conditional on a user visiting the article. Points are the maximum likelihood estimate of β_x^R, the effect of the feature on the article's value to the reader for each user type $x \in \{1, 2, 3\}$. Confidence intervals are the central 95% interval generated by 100 bootstrap replications, according to the method described in Section 4.