Newspaper Paywalls and Political Knowledge[*]

Julian Streyczek[†]

7 December, 2022

Preliminary and incomplete.
Please do not cite or circulate.

Abstract

I explore the impact of the widespread introduction of paywalls on US newspaper websites on readers' news consumption and political knowledge. Using newspaper-related Google searches as a proxy for online news consumption, I provide evidence that paywall launches reduce consumption of paywalled newspapers, especially for local newspapers. Furthermore, I show that regions where a larger fraction of online news consumption was affected by paywalls exhibit lower knowledge of local politics, as measured by knowledge of the majority party in State Senate and State House.

[*] I express deep gratitude to my advisors Carlo Schwarz and Guido Tabellini for their outstanding support. I also thank Alberto Manconi, Enrico Stivella, Eugen Dimant, Felix Gerding, Jaime Marques Pereira, Jan Bakker, Julia Cagé, Konstantin Sommer, Malin Kimmich, and Sarah Eichmeyer for valuable discussions and feedback. All errors remain my own. I thank Bocconi University for funding this chapter of my research journey.

[†] Università Bocconi, Department of Economics. julian.streyczek@phd.unibocconi.it

1 Introduction

In the past 15 years, paywalls have transformed much of US newspapers' online content from a public good into a club good: While throughout the first decade of the century most newspapers granted free access to articles on their websites, by 2019 around 75 percent of the top 25 US newspapers required visitors to pay a monthly fee for full access (Simon and Graves, 2019). Although paywalls help newspapers to stay profitable in times of declining advertising revenues, they also constitute barriers to information that have caused considerable reduction in web traffic to affected websites (Kim et al., 2020; Chiou and Tucker, 2013; Pattabhiramaiah et al., 2019).

Access to newspaper content can affect voter knowledge and participation, which can propagate to accountability of elected representatives and allocation of public funds (Gentzkow et al., 2011; Drago et al., 2014; Cagé, 2020; Gao et al., 2020). While these papers usually study entry and exit of newspapers, paywalls have received little attention in the context of this literature. I attempt to fill this gap by estimating the effect of newspaper paywalls on online news consumption and political knowledge.

In my analysis, I use publicly available data on newspaper-related Google searches as a proxy for news consumption, which is inspired by Vosen and Schmidt (2011) and Choi and Varian (2012). These data are not only a feasible alternative to proprietary web traffic data, but they also allow measuring regional variation that is instrumental in identifying paywall effects. My analysis has two parts:

First, I verify that newspaper-specific Google searches drop after a newspaper introduces a paywall on its website. To control for contemporaneous trends such as the diffusion of the internet and social media, I exploit the sequential timing of paywall launches in a staggered difference-in-differences design that compares the most important US newspapers that introduced a paywall from 2011 onward. I find a sharp, robust, and persistent reduction of searches after paywall launches. The results are similar to those that other authors have previously obtained using web traffic data, suggesting that Google searches are a suitable proxy for news consumption.

Second, I show that regions with higher exposure to paywalls experience higher reduction in knowledge on local politics in a difference-in-differences design. I measure exposure to paywalls by approximating the regional share of online news consumption that was affected by paywalls. I measure regional political knowledge using questions in the nationally representative Cooperative Election Study (CES) that ask respondents to name the current majority party in their State Senate or State House. I find that in regions with above-median paywall exposure knowledge of majorities in *state* legislative bodies decreases by around 7 percentage points, while there is no effect on knowledge on majorities on the national level. This finding is consistent with the previous results that paywalls disproportionally affect consumption of local newspapers.

Related Literature

This paper contributes to several strands of literature. First, it relates to previous work on the effects of introducing paywalls on newspaper websites. Studying the *New York Times* paywall, Cook and Attari (2012) report survey evidence regarding consumer behavior and attitudes, while Pattabhiramaiah et al. (2019) analyze web traffic data. Chiou and Tucker (2013) conduct a field experiment by introducing paywalls for four local US-based newspapers. The closest paper to mine is Kim et al. (2020) who analyze web traffic of 42 US-based newspapers that introduced a paywall between 2010 and 2017. They find that paywalls reduce pageviews by 30% on average, while the effect is larger for local newspapers and varies by slant, content uniqueness, and topic composition. My paper studies a similar sample and time frame, but leverages regional variation in newspaper popularity to explore effects on political outcomes.

Second, my paper contributes to a mature literature on how the advent of new digital media induced consumers to switch away from media with high news content, which generally led to lower political knowledge and lower voter turnout. Notable papers study the diffusion of television in the late 20th century (Gentzkow, 2006; Durante et al., 2019) and high-speed internet in the early 21st century (Falck et al., 2014; Campante et al., 2018; Gavazza et al., 2019). To the best of my knowledge, I am the first to connect the advent of paywalls to this literature: While the arrival of new media increase opportunity costs of news consumption by shifting consumer preferences, the introduction of paywalls does so by raising monetary barriers to existing media. Furthermore, the above papers examine diffusion using instruments, while I exploit the sudden nature of paywall launches in the spirit of a natural experiment.

Third, this paper relates to the literature on the importance of voters' access to news for political knowledge, participation, and allocation of public funds (Besley and Burgess, 2002; Strömberg, 2004; Ferraz and Finan, 2008; Snyder Jr and Strömberg, 2010). The strand that is most closely related to my paper considers the political effects of entry and exit of newspapers (Gentzkow et al., 2011; Drago et al., 2014; Cagé, 2020; Gao et al., 2020). I propose analyzing paywalls as a recent and influential barrier to voter information that has not yet been studied in the context of this literature.

Finally, in a broader sense, my paper relates to the literature studying how newspapers react to fundamental changes in the newspaper market, one result of which are paywalls. Many researchers have looked in detail at the decline of advertising revenues and subsequent adjustments by newspapers in terms of staffing, pricing, topic composition, or slant (George and Waldfogel, 2006; Seamans and Zhu, 2014; Gentzkow et al., 2014; Angelucci and Cagé, 2019; Bhuller et al., 2020). Furthermore, some authors tie these transformations to voter turnout and party-line voting (Angelucci et al., 2020; Djourelova et al., 2021). While these papers examine responses of the supply side of news, my paper focuses in detail on reactions on the demand side. In light of the above evidence, however, a future version will have to account for the many possible changes that newspapers might implement alongside a paywall.

The remainder of this paper is organized as follows. Section 2 presents my sample of newspapers, provides background on paywalls, and introduces my main data sources. Section 3 estimates effects of paywalls on the paywall-introducing newspaper, and Section 4 estimates effects on political knowledge. Section 5 concludes.

2 Background and Data

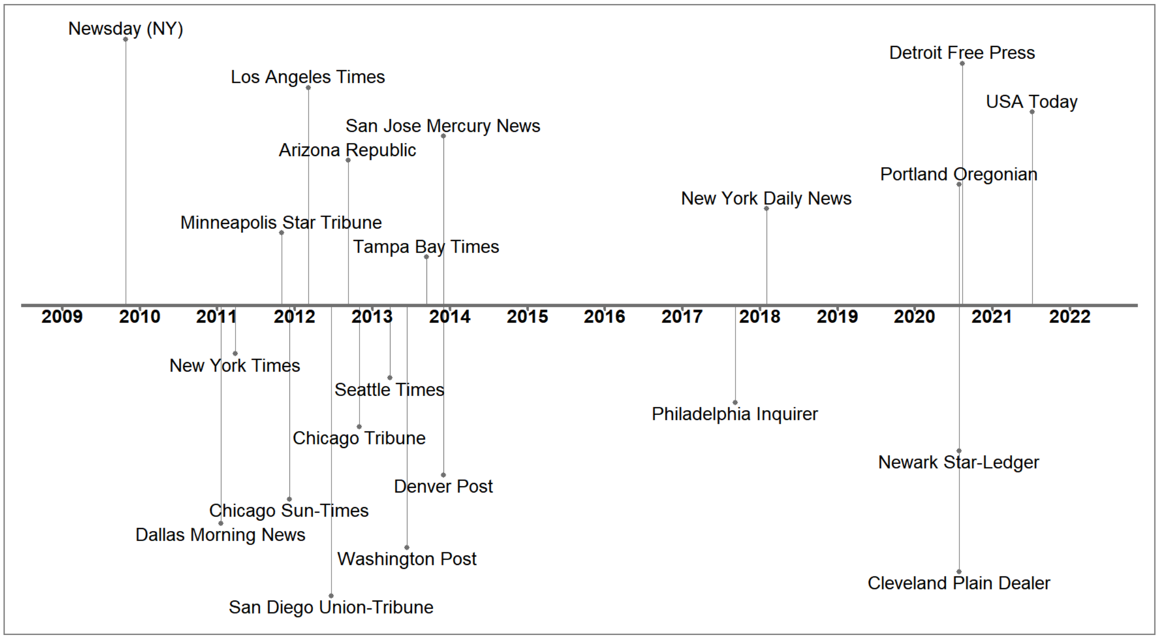
2.1 News outlets

My sample of newspapers comprises the largest 25 daily US newspapers by circulation in 2010 based on data from the Alliance for Audited Media (see Pew Research Center, 2011). I summarize the full list of newspapers in Table 1. This sample aims to cover the most influential paywall introductions on the US market. Unfortunately, due to the limited public availability of circulation data, it is currently not possible for me to include all relevant newspapers that introduced a paywall. However, given the large significance of the top 25 newspapers for the US news market, the current sample should be sufficient for the subsequent analyses.

2.2 Paywalls

I obtain information on paywall launches through news coverage by the paywall-introducing newspapers themselves or competing outlets. Where ambiguous, I check historical website snapshots for the appearance of a digital subscription button using the Wayback Machine. Among the 25 newspapers in my sample, one always had a paywall (Wall Street Journal), one never introduced one as of the writing of this paper (New York Post)[1], and two launched separate websites for premium content (Houston Chronicle, San Francisco Chronicle). For the remaining 21 newspapers, Figure 1 visualizes when the paywall was first introduced. Many newspapers launched their paywalls between 2011 and 2014, while others followed from around 2017.

Figure 1. Paywall launches of major US newspapers over time



Notes: Timing of first paywall on website among 25 top US newspapers by print circulation in 2010.

All but four newspapers introduced a ”metered” paywall that allows free access up to a certain number of articles.[2] A notable exception is Newsday which very early introduced a ”hard” paywall that fully restricted access to almost all content.[3] The other exceptions are the Cleveland Plain Dealer, Portland Oregonian, and Newark Star-Ledger (all owned by the publisher Advance Publications), which restricted access to selected ”premium” articles from July 2020. Moreover, paywalls are generally ”leaky” in the sense that they intentionally allow exceptions for users in private (”incognito”) browsing mode, or links that direct to specific articles from search engines or social media.[4]

[1] More accurately, the New York Post had a paywall in place for iPad users only, approximately between June 2011 and June 2012.

[2] The prominence of metered paywalls was partly due to Google’s ”First Click Free” policy, which until fall 2017 required news websites to grant each visitor at least 3 free articles per month to be listed on the Google search index.

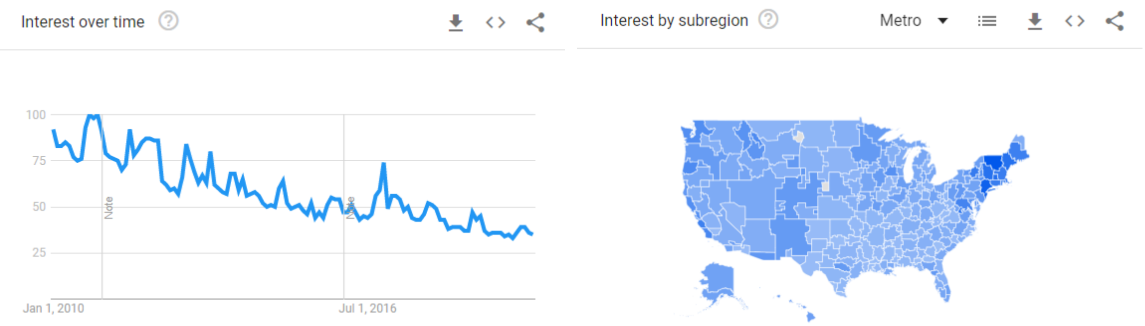
[3] However, free digital subscriptions were provided to customers of the parent company’s broadband service, which covered many of the newspaper’s core audience (Bercovici, 2011).

[4] In general, collecting and analyzing heterogeneity in paywall characteristics is difficult because newspapers can and do alter them flexibly, while producing little record of these practices.

2.3 Measuring interest using Google searches

One key data source for this paper is Google Trends, a public website maintained by Google that reveals information on the popularity of queries on Google search. For a user-specified keyword, location, and time frame, the service returns an index that captures how many among all conducted searches included the keyword. This index is scaled between 0 and 100, where 100 represents the highest relative search frequency in the time series. Search interest is reported across time (for a given region) and across regions (for a given time frame), as visualized exemplary in Figure 2, where region refers to Designated Market Area (DMA).[5] When entering multiple (up to 5) keywords, the relative search frequencies are scaled accordingly.

Figure 2. Information on searches as provided by Google Trends



(a) Timeline

(b) Map

Notes: Screenshots of output provided by Google Trends for keyword `"New York Times"` and time frame `2010-01-01` to `2019-12-31`. Left panel: Search interest in DMA ”New York NY” over time. Right panel: Average search interest across 10 years by DMA.

For each news outlet I identify a set of keywords that direct Google search users to the outlet or its website, such as `new york times`, `nytimes` and `ny times` for the New York Times. For details on keyword selection see Appendix A.1. Using these keywords, I scrape Google Trends for each newspaper and DMA between January 2005 and May 2022. Since I obtain information on searches across both time and regions, I combine the results to a panel of search interest. I outline details on the construction of the panel in Appendix A.2.

The panel covers 23 newspapers across 208 DMAs and 209 months, containing a total of 999,856 observations. Table 1 reports summary statistics. The variable *searches* exhibits considerable positive skewness. In fact, for most newspapers, searches are concentrated in relatively few regions.[6]

[5] DMAs are geographical areas in the US defined by the media analytics firm Nielsen.

[6] For a detailed discussion on the peculiarities of Google search data see Appendix A.3.

Table 1. Summary Statistics for Google searches

Newspaper	**Mean**	**SD**	**q50**	**q75**	**q95**	**Min**	**Max**	**N**
Arizona Republic	0.47	4.46	0.00	0.00	0.75	0	100	43,472
Chicago Sun-Times	0.80	4.21	0.00	0.00	3.23	0	100	43,472
Chicago Tribune	1.28	3.72	0.36	1.14	4.85	0	100	43,472
Detroit Free Press	0.33	1.73	0.00	0.12	0.72	0	100	43,472
Dallas Morning News	1.32	4.94	0.00	0.68	5.79	0	100	43,472
Denver Post	0.33	2.77	0.00	0.00	0.43	0	100	43,472
Los Angeles Times	1.71	5.57	0.00	1.27	8.13	0	100	43,472
San Jose Mercury News	0.66	4.45	0.00	0.00	1.75	0	100	43,472
Newsday (NY)	1.42	5.05	0.00	0.80	6.73	0	100	43,472
New York Daily News	0.63	2.34	0.00	0.00	3.97	0	100	43,472
New York Post	3.47	7.27	1.53	3.62	14.10	0	100	43,472
New York Times	7.89	7.85	6.34	11.12	22.32	0	100	43,472
Philadelphia Inquirer	0.52	4.13	0.00	0.00	1.28	0	100	43,472
San Diego Union-Tribune	0.32	3.17	0.00	0.00	0.74	0	100	43,472
Seattle Times	0.90	4.85	0.00	0.41	2.91	0	100	43,472
Minneapolis Star Tribune	1.18	6.54	0.00	0.00	2.20	0	100	43,472
Tampa Bay Times	0.37	4.23	0.00	0.00	0.46	0	100	43,472
Portland Oregonian	1.00	5.96	0.00	0.00	1.89	0	100	43,472
Cleveland Plain Dealer	0.68	4.66	0.00	0.00	1.96	0	100	43,472
Newark Star-Ledger	0.73	4.09	0.00	0.00	3.46	0	100	43,472
USA Today	6.22	4.45	5.42	8.11	13.73	0	100	43,472
Washington Post	4.45	6.03	3.51	5.71	12.21	0	100	43,472
Wall Street Journal	5.55	4.87	4.77	7.99	14.02	0	100	43,472
All	**1.84**	**5.31**	**0.00**	**1.06**	**9.51**	**0**	**100**	**999,856**

Notes: Summary statistics by newspaper for outlet-month-DMA-specific index for Google searches. The columns q50, q75, an q95 denote the 50th, 75th, and 95th quantile, respectively.

2.4 Political Knowledge

I measure political knowledge using a set of questions in the Cooperative Congressional Election Study (CES). The CES is a yearly, nationally representative survey of the US population administered by YouGov that uses repeated cross sections of 20,000 to 50,000+ respondents across the years 2006–2020. I focus on the following type of question:

> *Which party has a majority of seats in the [US Senate]?*

This question is asked separately for 4 legislative bodies: The US Senate and US House (in all years), and the upper and lower chamber (Senate and House) of the respondent's state (in all years except 2006 and 2009).

Table 2 presents summary statistics regarding the fraction of correct answers. Around 70 percent of respondents correctly name the majority in national Senate and House, while the number is around 50 percent for the respondent's respective State Senate and State House. The survey includes information on each respondent's county of residence, which allows measuring average political knowledge by region. Moreover, the survey includes self-reported information on respondents' demographics and other characteristics such as party identification or news interest.

Table 2. Summary statistics for variables derived from CES survey

Variable	**Mean**	**SD**	**Min**	**Max**	**N**
Knows majority: US Senate	0.63	0.48	0	1	521,149
Knows majority: US House	0.66	0.47	0	1	524,363
Knows majority: State Senate	0.48	0.50	0	1	475,817
Knows majority: State House	0.48	0.50	0	1	475,555
Age	46.93	17.00	18	109	557,455
Female	0.52	0.50	0	1	531,755
White	0.72	0.45	0	1	557,455
College	0.37	0.48	0	1	557,455
Family Income > 60k	0.35	0.48	0	1	557,455
Family Income > 100k	0.15	0.36	0	1	557,455
Employed	0.46	0.50	0	1	557,455
Unemployed	0.22	0.41	0	1	557,455
Retired	0.19	0.39	0	1	547,163
Democrat (pid3)	0.38	0.49	0	1	506,989
Independent (pid3)	0.32	0.47	0	1	506,989
Republican (pid3)	0.30	0.46	0	1	506,989
Liberal (ideo5)	0.24	0.42	0	1	557,455
Moderate (ideo5)	0.32	0.47	0	1	557,455
Conservative (ideo5)	0.34	0.47	0	1	557,455
News interest: Some	0.39	0.49	0	1	557,455
News interest: Much	0.45	0.50	0	1	557,455

Notes: Summary statistics of survey respondent–level variables derived from CES survey, averaged across all years. Indicator for party identification derived from 3-point scale. Indicator for ideology derived from 5-point scale, where "Very Liberal" and "Liberal" are coded as the latter, for example. News interest derived from 4-point scale, where "Much" is highest and "Some" is third-highest. Survey weights included.

3 Effect of paywalls on news consumption

3.1 Empirical strategy

In this section I analyze how paywall introductions on major US newspaper websites affected consumption of these newspapers online. I employ a staggered difference-in-differences design that exploits the sequential timing of paywall introductions to separate treatment effects from confounding trends. Specifically, my analysis is based on the following two-way fixed effects (TWFE) regression:

$$\log(\mathit{searches}_{nrt}) = \alpha_{nr} + \gamma_t + \beta\, \mathit{Paywall}_{nt} + \varepsilon_{nrt} \ , \tag{1}$$

where $\mathit{searches}_{nrt}$ denotes search index for newspaper n in region (DMA) r, at time (year-month) t. I choose newspaper-region as the unit of observation, which leverages the granularity of my data to control for potential regional confounders. In particular, α_{nr} denotes a unit fixed effect capturing several sources of time-constant heterogeneity, such as (the constant part of) demographics, newspaper characteristics, or the importance of a newspaper in a DMA. More subtly, α_{nr} addresses heterogeneous mapping from search index to news consumption across newspapers and regions. The time fixed effect γ_t captures common time trends, most notably long-term trends such as the general decrease of newspaper Google searches (compared to all searches), and short-term trends such as events of high news interest like elections or shootings. I choose a log-linear specification because the distribution of *searches* exhibits positive skewness. $\mathit{Paywall}_{nt}$ is a dummy that switches to one in the month t that newspaper n introduces a paywall on its website.

My sample comprises all newspapers that have introduced a paywall as of the writing of this paper. I use the years 2005 through 2019, so that i) the control group is always large enough, and ii) all treated units have a long enough post-treatment period. Specifically, the treated units are 15 newspapers that introduced their paywall between March 2011 and February 2018, while the pure control units are 5 newspapers that launched their paywall from 2020 onward. In my preferred specification, I additionally exclude 3 newspapers: Newsday, with an early and unusual paywall model; the Dallas Morning News, which introduced its paywall in the month that Google changed how it matches searches to geographical regions;[7] and the Washington Post, which was acquired by Jeff Bezos six months after its paywall launch

[7] See the hint in time series figures on Google Trends in January 2011.

and subsequently experienced major changes in its business model (Marx and Clark, 2014). I sensitivity of my results to these choices in several robustness checks. The coefficient β captures the average treatment effect on the treated (ATT) of earlier paywalls launched by major US newspapers; specifically, the approximate percentage effect of a paywall on newspaper-specific searches, averaged across newspapers and regions.

Clearly, one potential concern is the endogeneity of paywalls: If certain newspaper characteristics correlate with paywall introductions and long-term trends in newspaper popularity, the estimates might be biased. While for now I cannot fully rule out this concern, I take measures to alleviate it: First, I construct my sample as homogeneous as possible by comparing treated to later-treated units (as opposed to never-treated), to mitigate concerns of selection into treatment based on unobserved characteristics. Nevertheless, I will check the sensitivity of my results against alternative definitions of treatment and control groups. Second, if newspapers selected into launching paywalls earlier according to differing popularity trends, this confounder should materialize in negative pre-treatment trends, which I check.

Another concern regards the stable unit treatment value assumption (SUTVA), which requires no spillovers between units of analysis. This assumption may be violated, since a paywall on one news website might induce consumers to switch to a substitute outlet. Therefore, one should keep in mind that the estimates might contain a small downward bias.

Finally, many researchers have recently pointed out that the standard (naive) OLS estimator for TWFE regressions in settings with staggered treatment is likely biased if treatment effects are heterogeneous across units or time (for an overview see Baker et al., 2022). The reason is that the the OLS TWFE estimator for the ATT β weights comparisons of treated units to never-treated, not-yet-treated, and earlier-treated units (Goodman-Bacon, 2021). The latter comparison is problematic because treatment effect heterogeneity can lead to violation of the assumption of parallel trends between treated and earlier-treated units. Therefore, I will report results obtained from the robust estimator by Sun and Abraham (2021). This estimator uses as comparison units only the last-treated and never-treated units to estimate separate ATTs for each treatment cohort and period, which are then weighted by cohort sizes to obtain a consistent estimator for the ATT β.

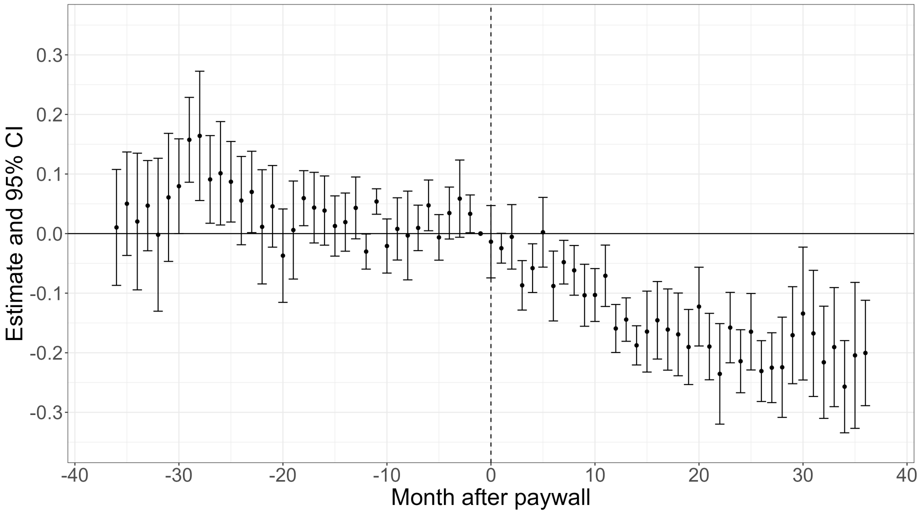
3.2 Results

Main results

Figure 3 shows estimates and 95 percent confidence intervals for period-wise paywall effects obtained using the Sun and Abraham (2021) estimator. After introducing a paywall, newspapers experience a steady reduction in popularity as measured by Google searches, which stabilizes after around 12 months. There seem to be little to no pre-treatment trends, indicating that selection into treatment according to popularity trends is not a major issue. Columns (3) and (4) in Table 3 report the corresponding pre-post estimates, suggesting that on average across newspapers and regions, paywalls reduce newspaper popularity by around 17 percent.

The results are qualitatively identical to those obtained from the naive (OLS) estimator, although smaller in size (see Figure B1 and columns (1) and (2) in Table 3 in the appendix). Moreover, the pattern is robust to i) including the 3 previously excluded paywalled newspapers, ii) Including the Wall Street Journal and New York Post as never-treated in the control group, iii) excluding 2 newspapers which dropped their paywall at some point, iv) using only years 2008 through 2016, which moves 2 more newspapers to the pure control group, and v) balancing the panel (see Table B1 for details).

Figure 3. Effect of paywalls on newspaper searches



Notes: Point estimates and 95% confidence intervals from a regression of (log) newspaper-related Google searches on month relative to paywall introduction, controlling for newspaper-DMA and month fixed effects. Estimated using robust Sun and Abraham (2021) estimator. Standard errors clustered by newspaper.

Table 3. Effect of paywalls on newspaper searches (Equation 1)

Dependent Variable:	log(searches)			
Estimator:	OLS		SA21	
Model:	(1)	(2)	(3)	(4)
Paywall	-0.2609***	-0.2655**	-0.1718***	-0.1689**
	(0.0835)	(0.0972)	(0.0556)	(0.0602)
DMA-newspaper FE	✓	✓	✓	✓
Month FE	✓		✓	
DMA-month FE		✓		✓
Observations	168,457	168,457	168,457	168,457
R^2	0.94	0.96	0.95	0.96

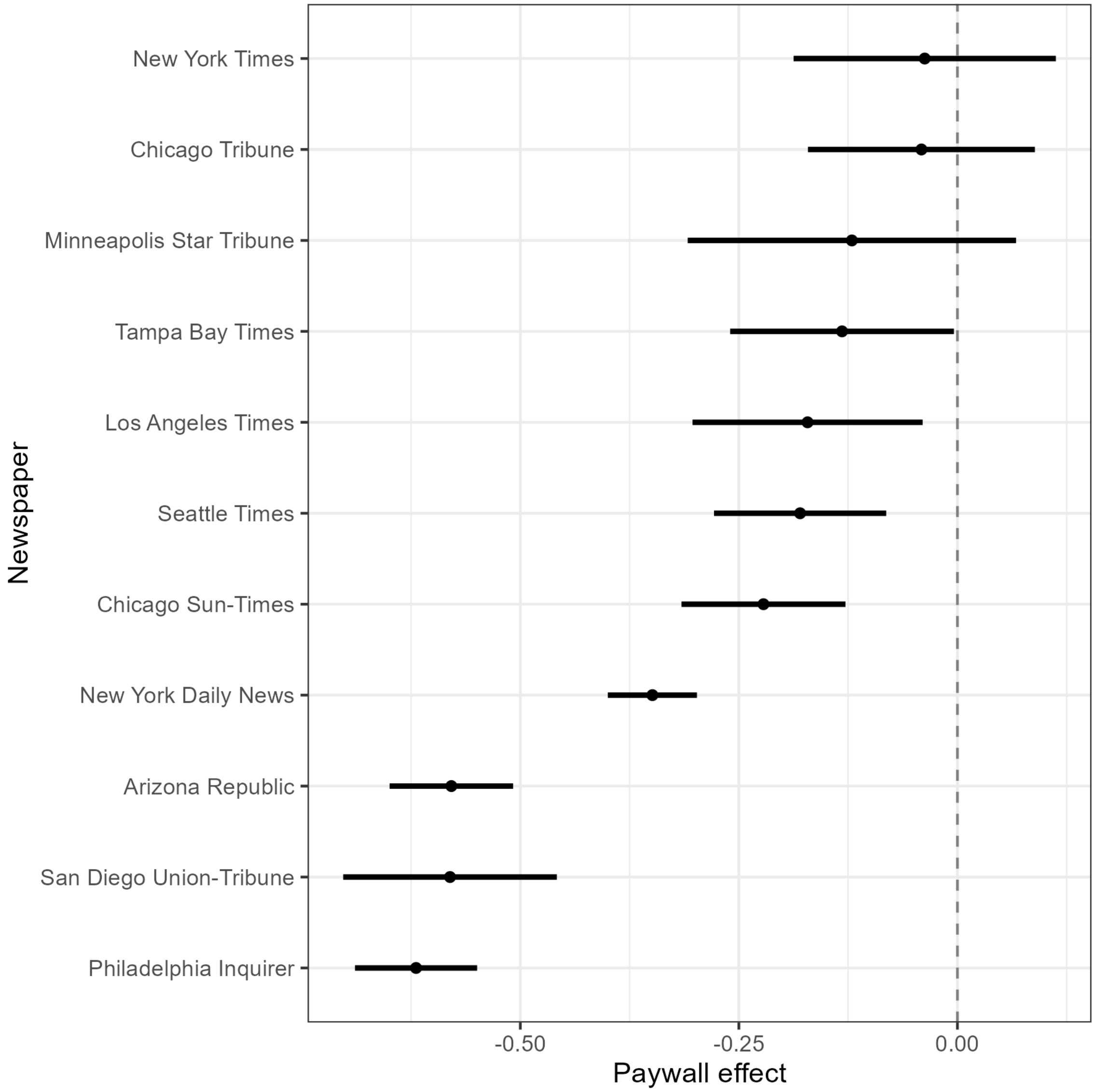
Notes: Columns (1) and (2) estimated using OLS, columns (3) and (4) estimated using robust Sun and Abraham (2021) estimator. Standard errors clustered by newspaper. Significance levels: *: 0.1, **: 0.05, ***: 0.01.

Heterogeneity by newspaper

Figure 4 visualizes paywall effect estimates by newspaper. Estimates vary greatly and range between 0 and -63 percent, with around two thirds significant. Notably, the effect for the two largest newspapers, the New York Times and the Los Angeles Times, seems to be rather small, while the largest effects are experienced by the smallest newspapers.

These results are consistent with Kim et al. (2020) who estimate paywall effects for 42 newspapers on web traffic. Their average estimated effect is around twice as high as mine at 30 percent, which is likely due to their larger share of local newspapers, for which also they find larger paywall effects. Unfortunately, the authors do not report newspaper-specific estimates for most newspapers, which makes a more detailed comparison difficult. Still, the similarity of my results obtained using Google searches to results obtained via actual news consumption data establishes further confidence that Google searches are a suitable proxy.

Figure 4. Effect of paywalls on newspaper searches, by newspaper



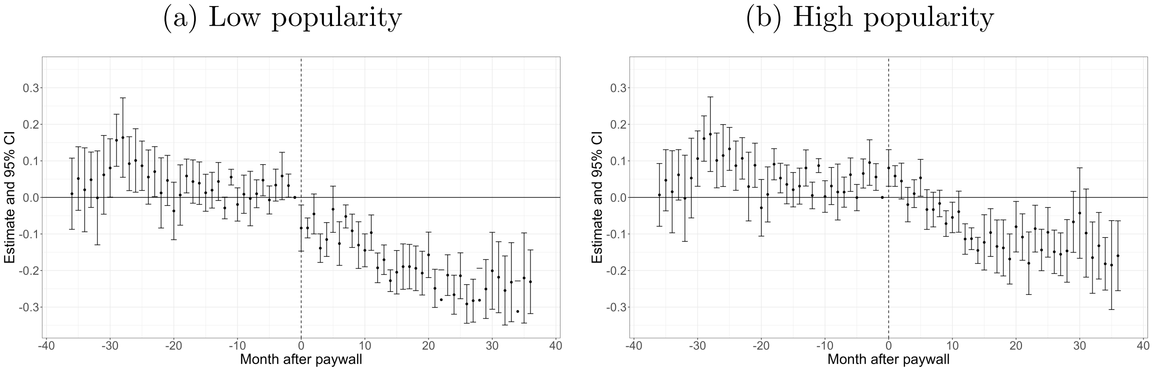
Notes: Point estimates and 95% confidence intervals from a regression of (log) newspaper-related Google searches on month relative to paywall introduction, controlling for newspaper-DMA and month fixed effects. Estimated using robust Sun and Abraham (2021) estimator. Note that because the Sun and Abraham (2021) estimator computes estimates by cohort rather than unit of treatment, it is not possible to recover individual estimates for two newspapers that are omitted here: the San Jose Mercury News and the Denver Post. Standard errors clustered by newspaper.

Heterogeneity by newspaper popularity

Next, I explore heterogeneity in the effect of paywall introductions with respect to regional newspaper popularity. For reasons of data availability, I base the analysis on a measure for regional popularity *within* newspaper: First, I average each newspaper's *searches* by region across the 12 months before its paywall launch to get a stable measure for regional popularity that is unaffected by seasonal trends. Then, for each newspaper, I split regions into two groups at the within-newspaper median of this popularity measure. Finally, I estimate Equation (1) for two subsamples in either of which I drop among the treated observations in the post period one of the groups, respectively. This way, I estimate two separate ATTs for regions of high and low pre-paywall popularity.[8]

Figure 5 reports period-wise estimates that reveal considerable heterogeneity. While the negative paywall effect is pronounced and significant for both groups, it is almost twice as large in regions where a newspaper is less popular (see Table B2 in the appendix). This finding is consistent with the idea that newspapers generate higher utility in their core markets, so that consumers are more willing to give up consumption in areas where the newspaper is relatively less relevant. Moreover, the fact that fringe audiences disengage more from the paywall-introducing newspapers implies that paywalls may have a polarizing effect on news consumption.

Figure 5. Effect of paywall on newspaper searches, by regional newspaper popularity



(a) Low popularity

(b) High popularity

Notes: Point estimates and 95% confidence intervals from two regressions of (log) newspaper-related Google searches on month after paywall introduction, for subsamples with and without including within-newspaper above-median pre-paywall popularity observations in treatment group post-treatment. Includes newspaper-DMA and month fixed effects. Estimated using robust Sun and Abraham (2021) estimator. Standard errors clustered by newspaper.

[8] In principle, I could alternatively estimate heterogeneous treatment effects by adding interaction terms to Equation (1). However, estimating this model with the Sun and Abraham (2021) estimator is not straightforward, which is why I opt for the sample split.

4 Effect of paywalls on political knowledge

In this section I explore how paywalls affected public knowledge about party majorities in US legislative bodies. I do so by exploiting variation across regions in exposure to paywalls. If trends in knowledge are similar across regions of high and low paywall exposure, then comparing these regions allows identifying a causal paywall effect.

4.1 Measuring exposure to paywalls

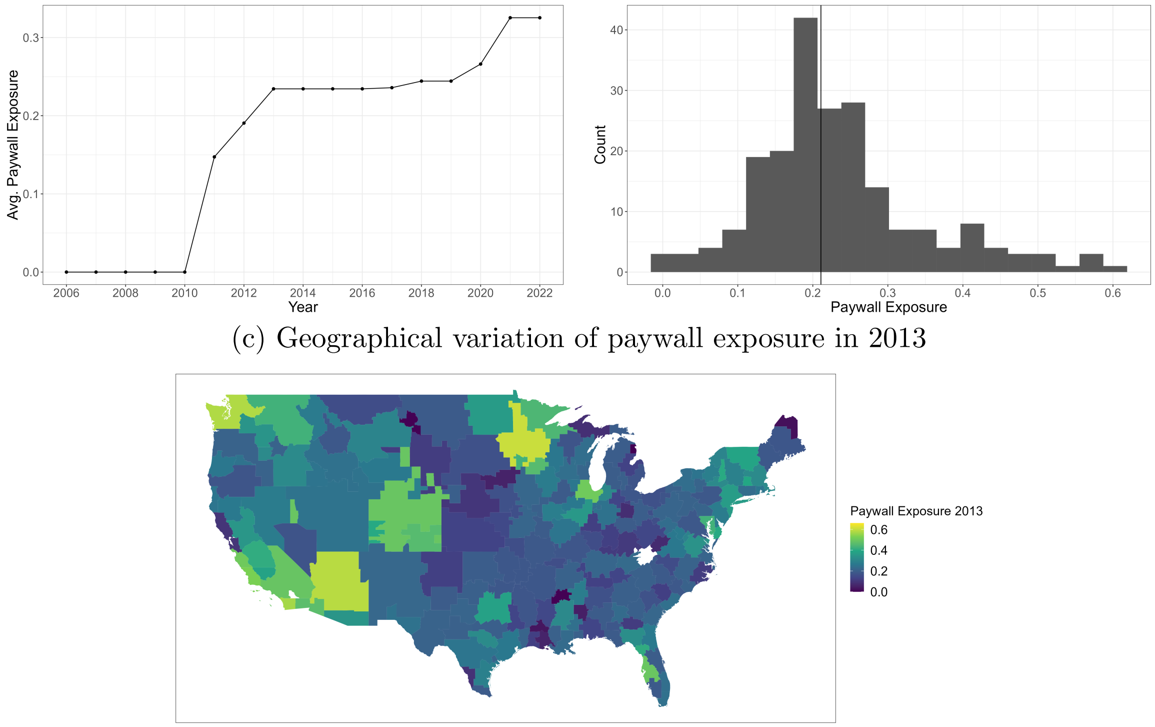
I measure regional paywall exposure by approximating how much online news consumption has been affected by paywalls. As in the previous section, I use Google searches as a proxy for news consumption.[9] I construct the measure as follows: First, for each region, I collect the share of searches directing to each of the newspapers in my sample in the year 2010, just before the first wave of paywalls. Second, for each year, I compute the share of searches directing to newspapers that have introduced a paywall. The idea is similar to the shift-share (”Bartik”) instrument, with relative pre-paywall consumption representing the ”share”, and the introduction of a paywall characterizing the ”shift”. Formally, for DMA r and year t I define

$$\textit{Paywall Exposure}_{\,r,t} = \frac{\sum_{n \,\in\, \text{paywalled newspapers}_{\,t}} \text{searches for } n_{\;r,\,2010}}{\sum_{m \,\in\, \text{all newspapers}} \text{searches for } m_{\;r,\,2010}} \,. \tag{2}$$

Figure 6a plots average *Paywall Exposure* across DMAs by year. Exposure to paywalls increases sharply between 2011 through 2013 and remains mostly flat afterwards. Figure 6b shows the distribution of *Paywall Exposure* at the end of 2013 across DMAs. The variable exhibits considerable variation, with some regions not being affected at all, to other regions where around 60 percent of pre-paywall online news consumption directed to newspapers that introduced a paywall soon after. Figure 6c visualizes the geographical variation of *Paywall Exposure* at the end of 2013, suggesting that paywalls were not concentrated in particular regions.

[9] For a technical note on the data collection for this step of the analysis see Appendix A.4.

Figure 6. Timeline and distribution of paywall exposure



(a) Average paywall exposure by year

(b) Distribution of paywall exposure in 2013

(c) Geographical variation of paywall exposure in 2013

Notes: Descriptive information on paywall exposure measure derived from Google search interest. Panel (a) shows average exposure across DMAs by year. Panel (b) shows the distribution of paywall exposure over DMAs in 2013 with the vertical line indicating the median. Panel (c) shows the geographical distribution of paywall exposure in 2013 over DMAs.

4.2 Empirical strategy

I exploit the sharp increase of paywall exposure between 2011 through 2013 in a difference-in-differences design. The treatment variable is a binary variable indicating regions with "high" paywall exposure. In particular, the binary variable $High\ PWX$ equals one for DMAs with above-median $Paywall\ Exposure$ in 2013.

The use of a binary treatment variable circumvents potential biases that arise in difference-in-differences regressions when treatment is heterogeneous across units and time (De Chaisemartin and d'Haultfoeuille, 2020; Callaway and Sant'Anna, 2021). Such issues are typically aggravated when the treatment variable is continuous because different "dosage" of treatment can have different marginal effects (Callaway et al., 2021; de Chaisemartin et al., 2022).[10]

[10] In particular, if treatment is continuous, the target parameter of the estimation is the average causal response of the treatment as a *function* of treatment dose d.

I estimate the following model:

$$Knowledge_{\,i,r,t} = \alpha_r + \gamma_t + \beta\, I\big(High\ PWX_{\,r}\big) \times I(t \geq 2011) + \delta\, \mathbf{X}_{i,r,t} + \varepsilon_{i,r,t}\ . \tag{3}$$

The dependent variable $Knowledge_{\,i,r,t}$ is a dummy equal to one if respondent i in county r and year t correctly names the majority party in a given legislative body. The county fixed effect α_r accounts for pre-existing regional differences in political knowledge, and the year fixed effect $gamma_t$ captures any national changes in knowledge. $I(High\ PWX_{\,r})$ denotes a dummy equal to one for regions with above-median paywall exposure in 2013. $\mathbf{X}_{i,r,t}$ denotes individual control variables for each survey respondent (age, sex, education, income, party identification, ideology).

Under the assumption that knowledge in low-exposure and high-exposure regions would have evolved the same in the absence of paywalls, β identifies the causal effect of high paywall exposure on knowledge in affected regions.

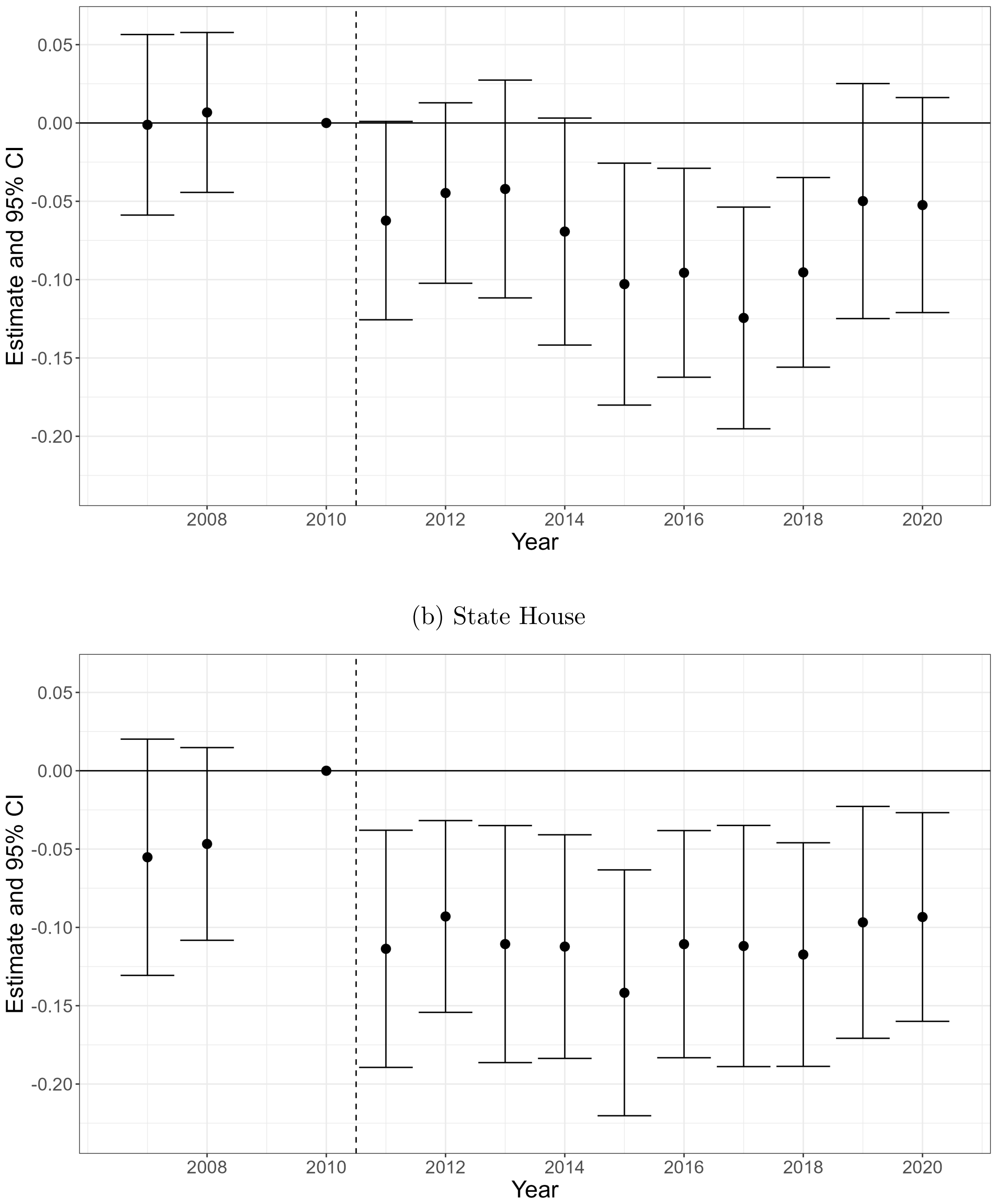
4.3 Results

Main results

I estimate regressions in the form of Equation (3) using OLS. The two panels in Figure 7 show yearly treatment effect estimates for knowledge on majorities in *State* Senate and *State* House, respectively. The estimates suggest no differential trends in knowledge before paywalls. Furthermore, there is a sharp and persistent drop in knowledge in regions in which online news consumption was likely affected relatively much by paywalls.

Columns (1) and (2) of Table 4 report corresponding pre-post estimates. In regions with high paywall exposure, knowledge about *State* Senate and *State* House majorities decreased by around 7 percentage points from an average of around 50 percent each. Columns (3) and (4) reveal that there is no statistically significant effect on knowledge on *national* majorities. This finding is consistent with the previous result that paywalls mainly reduced online consumption of local newspapers, which should disproportionally affect knowledge on local politics.

Figure 7. Effect of paywalls on knowledge on party majority in state legislative bodies



(a) State Senate

(b) State House

Notes: Point estimates and 95 percent confidence intervals for a regression of knowledge on the party majority in respondents' state Upper / Lower Chamber on the interaction of high regional paywall exposure in 2013 and a yearly dummy, controlling for respondent characteristics, county fixed effects, and year fixed effects. Survey weights are included. Standard errors clustered by DMA.

Table 4. Effect of paywall exposure on political knowledge

Dependent Variable:	Knowledge of Majority			
Level:	State		National	
Chamber:	Senate	House	Senate	House
Model:	(1)	(2)	(3)	(4)
High Paywall Exposure	-0.0757***	-0.0763***	-0.0078	-0.0090
	(0.0211)	(0.0195)	(0.0067)	(0.0065)
County FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓
Observations	417,071	416,863	456,252	459,115
R^2	0.18	0.17	0.19	0.20

Notes: The dependent variable in columns (1) and (2) is a dummy equal to one if the survey respondent correctly names the party majority in their state's upper and lower chamber, respectively. Columns (3) and (4) concern knowledge of the national US Senate and House, respectively. All regressions include individual controls and survey weights. Standard errors clustered by DMA. Significance levels: *: 0.1, **: 0.05, ***: 0.01.

Robustness checks

Next, I verify that the magnitude of the effect on knowledge increases with paywall exposure. I re-run the baseline regression in Equation (3) replacing the dummy for above-median paywall exposure in 2013 with a discrete variable indicating terciles and quartiles, respectively. Table 5 shows the results. For knowledge on both State Senate and State House, the point estimates increase with the quartile of paywall exposure. The large coefficient on the fourth quartile reveals that much of the effect is driven by regions where paywalls affected a large share of online news consumption.

I also check that my results are not driven by changes in state majorities that coincide with paywalls. If some individuals update information slowly, then the paywall effect estimate for regions with more majority changes might be biased downwards. I add control variables that indicate whether the majority in the State Senate or State House changed, respectively, during the year of the survey. Table 6 reports results. The coefficients on the two control variables are highly significant, suggesting delayed information updating among some individuals. The paywall effect estimates decrease only slightly, indicating that changes in state majorities are not the driving factor behind the average reduction in knowledge.

Table 5. Heterogeneous effect of paywall exposure by quantiles

Dependent Variable:	Knowledge of majority in state legislation					
Chamber:	Senate			House		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Above-median Paywall Exposure	-0.0757***			-0.0763***		
	(0.0211)			(0.0195)		
2nd tercile of Paywall Exposure		-0.0083			0.0171	
		(0.0274)			(0.0283)	
3rd tercile of Paywall Exposure		-0.0918***			-0.0794***	
		(0.0252)			(0.0211)	
2nd quartile of Paywall Exposure			-0.0333			0.0085
			(0.0304)			(0.0312)
3rd quartile of Paywall Exposure			-0.0499*			-0.0259
			(0.0255)			(0.0302)
4th quartile of Paywall Exposure			-0.1225***			-0.1003***
			(0.0268)			(0.0250)
County FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓
Observations	417,071	417,071	417,071	416,863	416,863	416,863
R^2	0.18	0.18	0.18	0.17	0.17	0.17

Notes: Regressions similar to Equation (3), replacing the dummy for above-median paywall exposure in 2013 with a discrete variable indicating terciles and quartiles, respectively. All regressions include individual controls and survey weights. Standard errors clustered by DMA. Significance levels: *: 0.1, **: 0.05, ***: 0.01.

Table 6. Effect of paywall exposure after controlling for recent majority changes

Dependent Var.:	Knowledge of majority in state legislation					
Chamber:	Senate			House		
Model:	(1)	(2)	(3)	(4)	(5)	(6)
High PWX	-0.0757*** (0.0211)	-0.0750*** (0.0203)	-0.0682*** (0.0195)	-0.0763*** (0.0195)	-0.0542*** (0.0141)	-0.0536*** (0.0142)
St. Senate Change		-0.0730** (0.0331)	-0.0585* (0.0309)			0.0394** (0.0180)
St. House Change			-0.0670*** (0.0176)		-0.2144*** (0.0155)	-0.2241*** (0.0134)
County FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓
Observations	417,071	417,071	417,071	416,863	416,863	416,863
R^2	0.18	0.18	0.18	0.17	0.18	0.18

Notes: Regressions similar to Equation (3), adding control variables indicating whether the majority in the State Senate or State House changed, respectively, during the year of the survey. All regressions include individual controls and survey weights. Standard errors clustered by DMA. Significance levels: *: 0.1, **: 0.05, ***: 0.01.

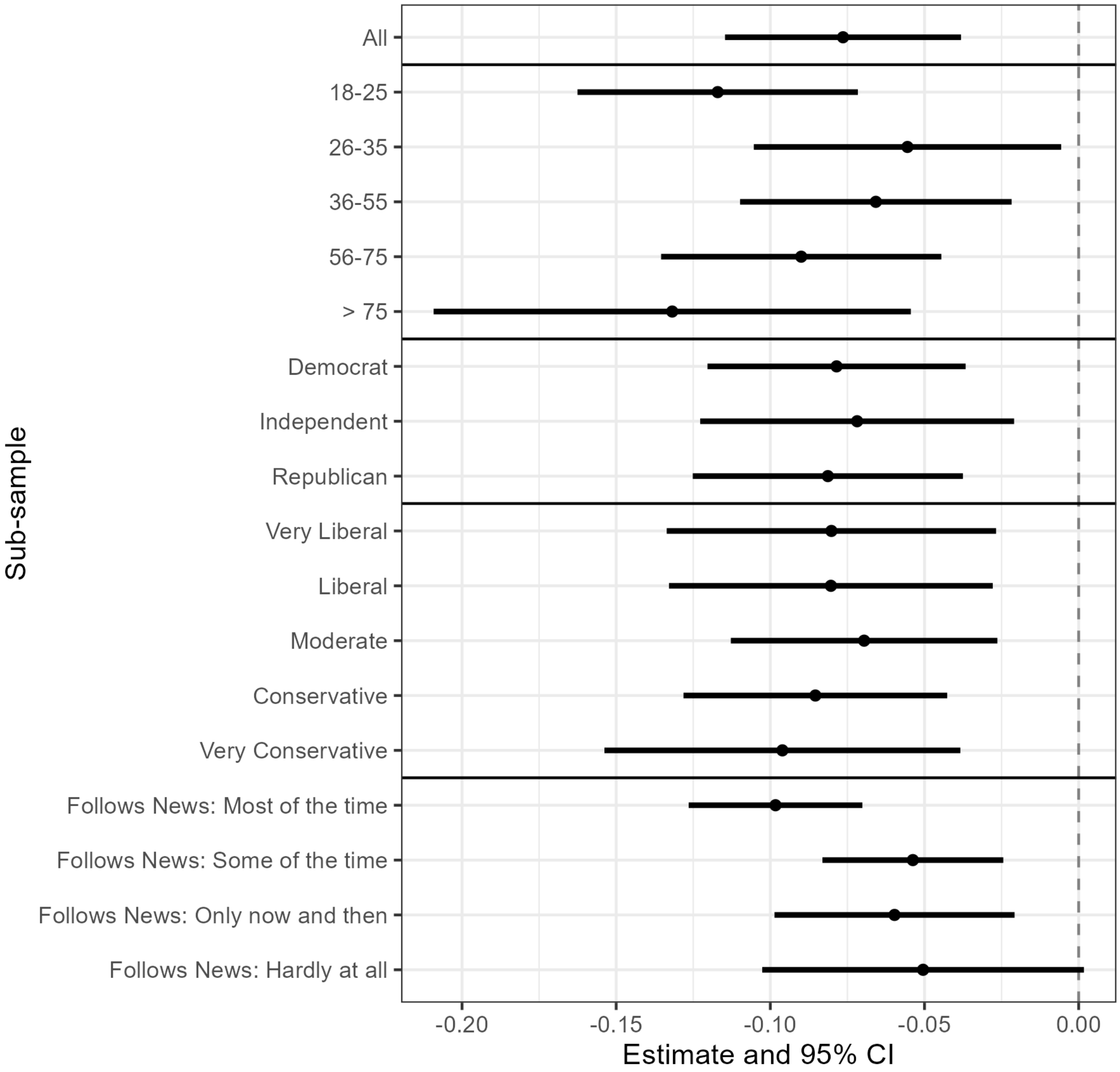
Heterogeneous effects

Finally, I demonstrate heterogeneity in the effect of paywall exposure. Figure 8 reports estimates for knowledge on State House majorities obtained by splitting the sample across groups of age, party identification, ideology, and news interest.

Although not all differences between groups are significant at the 5-percent level, the results are suggestive of the following patterns. First, the effect might be larger for young individuals between 18 and 25 years. This sub-population likely consumes a larger share of their news online and is therefore more exposed to paywalls. Moreover, these individuals grew up when the internet offered all kinds of information for free, which is why they could be especially reluctant to start paying for news. Second, the effect might be larger for republicans and conservatives. Since such individuals tend to value the concept of ”freedom” higher, they might be less willing to financially support newspapers that restrict free access to information. Third, the effect seems to be larger for individuals that report high news interest. This result is consistent with the idea that individuals with high news consumption are affected most by the reduction in free news.

The results for knowledge on State Senate majorities are qualitatively similar, but heterogeneity of effects is smaller. For more details, see Figure C2 in the appendix.

Figure 8. Heterogeneous effects of paywall exposure (State House)



Notes: Point estimates and 95 percent confidence intervals for coefficient on $High\ PWX$ on knowledge of State House majorities in Equation (3) in different subsamples. The first section (”All”) refers to the full set of observations. The second section refers to age. The third section refers to party identification. The fourth section refers to political ideology. The fifth section refers to the tendency of keeping up with the news. Survey weights are included. Standard errors clustered by DMA.

5 Conclusion

This paper links the appearance of paywalls on newspaper websites to changes in political knowledge by using newspaper-related Google searches as a proxy for online news consumption. First, I show that paywalls decrease consumption of affected newspapers, especially among local newspapers, which is in line with previous findings based on web traffic. Second, I show that survey respondents in regions where paywalls affected a large share of news consumption were less able to name the majority party in their state legislative bodies.

A future version of this paper could benefit from several extensions. First, I will address in more detail the potential endogeneity of paywall introductions. Ideally, I would instrument introduction, for example by leveraging regional differences in market forces that explain the timing of paywalls. Potential sources of exogenous variation could be the staggered introduction of Craigslist or the diffusion of social media, which might have reduced online advertising revenue of newspapers.

Second, I will explore substitution of readers between news outlets in response to paywalls. If outlets differ from the paywalled newspapers in topic composition, such patterns could provide additional mechanisms for the reduction of knowledge about local politics. Moreover, if the substitute outlets differ in political slant, paywalls might ultimately affect political polarization through persuasion.

Third, I will examine whether the reduction of knowledge affects participation in local elections. If internalized by politicians, such changes might ultimately affect politician effort, allocation of public funds, and policy implementation.

References

Angelucci, Charles and Julia Cagé, "Newspapers in times of low advertising revenues," *American Economic Journal: Microeconomics*, 2019, *11* (3), 319–64.

_ , _ , **and Michael Sinkinson**, "Media competition and news diets," Technical Report, National Bureau of Economic Research 2020.

Baker, Andrew C, David F Larcker, and Charles CY Wang, "How much should we trust staggered difference-in-differences estimates?," *Journal of Financial Economics*, 2022, *144* (2), 370–395.

Bercovici, Jeff, "More Proof That Paywalls Work From...Newsday?," *Forbes*, 2011. November 11. Available at: https://www.forbes.com/sites/jeffbercovici/2011/11/01/more-proof-that-paywalls-work-from-newsday. Last accessed 22 August 2022.

Besley, Timothy and Robin Burgess, "The political economy of government responsiveness: Theory and evidence from India," *The quarterly journal of economics*, 2002, *117* (4), 1415–1451.

Bhuller, Manudeep, Tarjei Havnes, Jeremy McCauley, and Magne Mogstad, "How the internet changed the market for print media," Technical Report, Memorandum 2020.

Cagé, Julia, "Media competition, information provision and political participation: Evidence from French local newspapers and elections, 1944–2014," *Journal of Public Economics*, 2020, *185*, 104077.

Callaway, Brantly and Pedro H.C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230. Themed Issue: Treatment Effect 1.

_ , **Andrew Goodman-Bacon, and Pedro HC Sant'Anna**, "Difference-in-differences with a continuous treatment," *arXiv preprint arXiv:2107.02637*, 2021.

Campante, Filipe, Ruben Durante, and Francesco Sobbrio, "Politics 2.0: The multifaceted effect of broadband internet on political participation," *Journal of the European Economic Association*, 2018, *16* (4), 1094–1136.

Chaisemartin, Clément De and Xavier d'Haultfoeuille, "Two-way fixed effects estimators with heterogeneous treatment effects," *American Economic Review*, 2020, *110* (9), 2964–96.

Chiou, Lesley and Catherine Tucker, "Paywalls and the demand for news," *Information Economics and Policy*, 2013, *25* (2), 61–69.

Choi, Hyunyoung and Hal Varian, "Predicting the present with Google Trends," *Economic record*, 2012, *88*, 2–9.

Cook, Jonathan E and Shahzeen Z Attari, "Paying for what was free: Lessons from the New York Times paywall," *Cyberpsychology, behavior, and social networking*, 2012, *15* (12), 682–687.

de Chaisemartin, Clément, Xavier D'Haultfoeuille, Félix Pasquier, and Gonzalo Vazquez-Bare, "Difference-in-Differences Estimators for Treatments Continuously Distributed at Every Period," *arXiv preprint arXiv:2201.06898*, 2022.

Djourelova, Milena, Ruben Durante, and Gregory Martin, "The impact of online competition on local newspapers: Evidence from the introduction of Craigslist," 2021.

Drago, Francesco, Tommaso Nannicini, and Francesco Sobbrio, "Meet the press: How voters and politicians respond to newspaper entry and exit," *American Economic Journal: Applied Economics*, 2014, *6* (3), 159–88.

Durante, Ruben, Paolo Pinotti, and Andrea Tesei, "The political legacy of entertainment TV," *American Economic Review*, 2019, *109* (7), 2497–2530.

Falck, Oliver, Robert Gold, and Stephan Heblich, "E-lections: Voting Behavior and the Internet," *American Economic Review*, 2014, *104* (7), 2238–65.

Ferraz, Claudio and Frederico Finan, "Exposing corrupt politicians: the effects of Brazil's publicly released audits on electoral outcomes," *The Quarterly journal of economics*, 2008, *123* (2), 703–745.

Gao, Pengjie, Chang Lee, and Dermot Murphy, "Financing dies in darkness? The impact of newspaper closures on public finance," *Journal of Financial Economics*, 2020, *135* (2), 445–467.

Gavazza, Alessandro, Mattia Nardotto, and Tommaso Valletti, "Internet and politics: Evidence from UK local elections and local government policies," *The Review of Economic Studies*, 2019, *86* (5), 2092–2135.

Gentzkow, Matthew, "Television and voter turnout," *The Quarterly Journal of Economics*, 2006, *121* (3), 931–972.

_ , Jesse M Shapiro, and Michael Sinkinson, "The effect of newspaper entry and exit on electoral politics," *American Economic Review*, 2011, *101* (7), 2980–3018.

_ , _ , and _ , "Competition and ideological diversity: Historical evidence from us newspapers," *American Economic Review*, 2014, *104* (10), 3073–3114.

George, Lisa M and Joel Waldfogel, "The New York Times and the market for local newspapers," *American Economic Review*, 2006, *96* (1), 435–447.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277.

Jr, James M Snyder and David Strömberg, "Press coverage and political accountability," *Journal of political Economy*, 2010, *118* (2), 355–408.

Kim, Ho, Reo Song, and Youngsoo Kim, "Newspapers' Content Policy and the Effect of Paywalls on Pageviews," *Journal of Interactive Marketing*, 2020, *49*, 54–69.

Marx, Greg and Anna Clark, "Can The Washington Post's national push help support local news?," *Columbia Journalism Review*, 2014. April 4. Available at: https://archives.cjr.org/united_states_project/washington_post_local_papers_partnership.php. Last accessed 11 August 2022.

Pattabhiramaiah, Adithya, S Sriram, and Puneet Manchanda, "Paywalls: Monetizing online content," *Journal of marketing*, 2019, *83* (2), 19–36.

Pew Research Center, "The State of the News Media 2011: An Annual Report on American Journalism," 2011. Available at: https://www.pewresearch.org/wp-content/uploads/sites/8/2017/05/State-of-the-News-Media-Report-2011-FINAL.pdf. Last accessed: 5 August 2022.

Seamans, Robert and Feng Zhu, "Responses to entry in multi-sided markets: The impact of Craigslist on local newspapers," *Management Science*, 2014, *60* (2), 476–493.

Simon, Felix M and Lucas Graves, "Pay models for online news in the US and Europe: 2019 Update," *Reuters Institute. https://reutersinsfitute. polifics. ox. ac. uk/sites/default/files/2019-05/Paymodels_for_Online_News_FINAL. pdf*, 2019.

Strömberg, David, "Radio's impact on public spending," *The Quarterly Journal of Economics*, 2004, *119* (1), 189–221.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Vosen, Simeon and Torsten Schmidt, "Forecasting private consumption: survey-based indicators vs. Google trends," *Journal of forecasting*, 2011, *30* (6), 565–578.

A Appendix: Data

A.1 Principled selection of keywords

Google Trends allows up to five keywords in a single query, which are then scaled with respect to each other by their relative popularity. For each newspaper, magazine, and news aggregator, I choose the keywords according to the following process, which is illustrated using ”The New York Times” as an example:

1. newspaper name (without ”the”, if applicable): `new york times`

2. newspaper name (without ”the”), with common abbreviation for location: `ny times`

3. center piece of URL hostname: `nytimes`, `nyt`

4. based on subjective considerations: exclude confounding terms: do not include `times`, add `-square` (excludes `new york times square`)

5. based on subjective considerations: exclude non-opinion-forming offers: add `-cooking`, `-crossword`, `-spelling`, `-bee`, `-wordle`

For TV and social media, I just use the outlet name, adding ”news” where ambigous (e.g. `fox news`). Table A1 shows the keywords I used for each outlet.

Alternatively, one could use Google's search ”topics”, which attempt to categorize searches in terms of the search subject. For example, this allows analyzing the search frequency for the company ”Apple” as opposed to the fruit. Therefore, for a predictor of the popularity of the New York Times, one could use the index for the *topic* `New York Times (Newspaper)` instead of the *keyword* `New York Times`. However, how topic indices are computed is essentially a black box, because Google releases almost no information on its methodology, which is why I opted for the more complicated but transparent approach above. In a future version of this paper, I will compare the performance of my current popularity indices to those derived from topic popularity.

Appendix Table A1. Keywords used to measure search interest of news outlets on Google search

	Type	Keyword 1	Keyword 2	Keyword 3	Keyword 4
Arizona Republic	Newspaper	arizona republic	az republic -arizona	az central -republic	azcentral
Chicago Sun-Times	Newspaper	chicago sun	chicago times -sun	suntimes	
Chicago Tribune	Newspaper	chicago tribune	chicagotribune		
Cleveland Plain Dealer	Newspaper	plain dealer	cleveland dealer -plain		
Dallas Morning News	Newspaper	dallas morning news	dallasnews	dallasmorningnews	
Denver Post	Newspaper	denver post -office -offices -hold -vacation	denverpost		
Detroit Free Press	Newspaper	detroit free press	freep	freep.com	
Los Angeles Times	Newspaper	los angeles times -crossword	la times -crossword	latimes -crossword	
Minneapolis Star Tribune	Newspaper	star tribune -casper -terre -haute -neighbor -chatham -ledger -indianapolis	minneapolis tribune -star	startribune -casper -terre -haute -neighbor -chatham -ledger -indianapolis	
New York Daily News	Newspaper	new york daily -mail -post -times -number	ny daily -mail -post -times -number	nydailynews + nydn	
New York Post	Newspaper	new york post -office -offices -hold -vacation	ny post -office -offices -hold -vacation	nypost	
New York Times	Newspaper	new york times -square -cooking -mini -crossword -spelling -bee -wordle	ny times -square -cooking -mini -crossword -spelling -bee -wordle	nytimes -cooking -mini -crossword -spelling -bee -wordle	nyt -cooking -mini -crossword -spelling -bee -wordle
Newark Star-Ledger	Newspaper	star ledger	star-ledger		
Newsday (NY)	Newspaper	newsday			
Philadelphia Inquirer	Newspaper	philadelphia inquirer -building	philly.com		
Portland Oregonian	Newspaper	oregonian	oregonlive		
San Diego Union-Tribune	Newspaper	san diego tribune	ut san diego	utsandiego	
San Jose Mercury News	Newspaper	mercury news	jose mercury -news	mercurynews	
Seattle Times	Newspaper	seattle times	seattletimes		
Tampa Bay Times	Newspaper	petersburg times	tampa times	tampabay	
USA Today	Newspaper	usa today	usatoday		
Wall Street Journal	Newspaper	wall street journal	wsj		
Washington Post	Newspaper	washington post -crossword -office -offices -hold -vacation	washingtonpost -crossword	wapo -crossword	

A.2 Constructing the panel of newspaper searches

For each news outlet, I scrape the relative search indices for all keywords by month and DMA between January 2005 and May 2022. Since Google Trends does not provide panel data, but either time series (for a given region) or cross sections (for a given time frame), I scrape monthly cross sections separately for each news outlet and scale them using the time series in the DMA with largest average search interest. Then, I add up the keyword-specific indices to obtain a single index that I call *searches*. By construction, this variable is scaled between 0 and 100 *within* each news outlet, where 100 indicates the DMA-month observation with highest relative search interest in that news outlet.

A.3 Peculiarities of Google search data

Google search data are a unique, inexpensive source of rich information on relative newspaper popularity over regions and time. However, Google Trends data exhibit some peculiarities that require particular care during the analysis. First, Google Trends rounds (interval-censors) its search index to integers between 0 and 100. Although this practice produces higher noise where/when newspapers are least popular, it usually implies no bias since the direction of the rounding error for a given observation is effectively random. The only exception is censoring at zero when considering log(*searches*), which may produce selection. Therefore, I cross-check my results in a balanced panel to verify that the results are not driven by observations entering and leaving the sample.

Second, Google Trends does not capture website visits via apps, browser bookmarks, links from other sources, or typing website URLs directly. Instead, the data capture users who search for coverage of a newspaper on specific topics (e.g. "NY Times Trump"), look for the URL, search for background information of the newspaper, or simply have a habit of accessing websites through Google. Therefore, one needs to keep in mind that the data are generated by a specific, non-representative sample from the US population. To address this concern, I will show that the information produced by this potentially specific sample reproduces findings based on web traffic, which is a more conventional measure of online content consumption.

Third, depending on newspaper, region, or both, the mapping from Google searches captured by my keyword selection to online news consumption can differ. For example, consider two outlets that differ in the uniqueness of their name. For the Wall

Street Journal, the keywords `wall street journal` and `wsj` will likely capture most of the searches towards its website or articles. For the Time magazine, in contrast, many users might use the keyword `time`, while fewer will use the combination `time magazine`. I only use the latter keyword in my analysis to avoid capturing searches for the clock time instead of the magazine, which implies that comparisons between different news outlets based on the *levels* of search indices are not useful. Moreover, this shortcoming may also be present when comparing searches for the same news outlet across different regions. For example, for the search query `times`, Google might refer a New York–based user to the New York Times, but a Los Angeles–based user to the Los Angeles Times. Because of this ambiguity, I do not use the keyword `times` at all, which likely overlooks more searches in regions where Google automatically proposes the link to the desired outlet. While these issues concern the *levels* of search indices, they should not affect *percentage changes* in search indices, which are rather driven by changes in search behavior. Thus, in my analysis I overcome these issues by comparing percentage changes in search interest over time, across news outlets and regions.

A.4 Collection and scaling of regional search interest

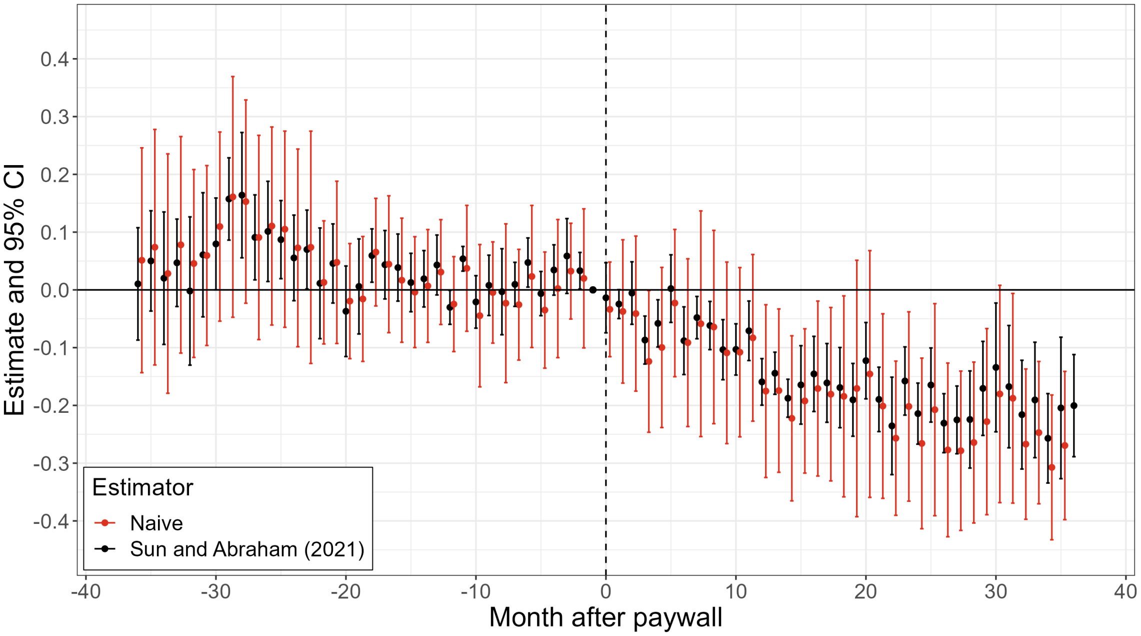
The construction of the variable *Paywall Exposure* requires scaling the search index for the different newspapers in my sample relative to each other. As mentioned in section 2, in the panel used in section 3 the searches of *each* newspaper are instead scaled between 0 and 100, and are therefore not meaningful for comparing search interest for different newspapers. I constructed the panel this way for practical reasons: First, the relative scaling does not affect the empirical strategy or results in the first because I am looking at differences in search interest within newspaper. Second, scaling search interest is tricky by design: i) Google Trends only allows comparisons of search interest for up to 5 keywords at a time, and ii) if popularity differs considerably between keywords, then search interest for keywords with smaller search interest is interval-censored due to rounding, which introduces noise.

Therefore, I re-collect data on newspaper-specific searches by region in 2010 using the Google Trends *topics* feature, using the fact that for each newspaper in my sample, Google defines a pre-existing topic that summarizes search interest for each newspaper, accessible via a single keyword.

Specifically, for each newspaper, I run a single query that includes the topic keyword for the newspaper and the topic keyword for ”news”, for searches spanning the year 2010, which provides me with a geographical breakdown of search interest. Since the search interest for the topic ”news” is (of course) identical for the same region across queries, and does not differ too much from the search interest of the most popular newspaper in a region, it serves as an ideal benchmark to correctly scale search interest of different newspapers within each region, for the year 2010.

B Appendix: Effect of paywalls on news consumption

Appendix Figure B1. Effect of paywall on newspaper searches, comparing naive TWFE estimator to Sun and Abraham (2021) estimator



Notes: Point estimates and 95% confidence intervals from a regression of (log) newspaper-related Google searches on month relative to paywall introduction, controlling for newspaper-DMA and month fixed effects. Standard errors clustered by newspaper-DMA.

Appendix Table B1. Effect of paywalls on newspaper searches (Equation 1)

Dependent Variable:	log(searches)					
	Baseline	Balanced	All pw'd	Non-pw'd	No dropped pw	Large control
	(1)	(2)	(3)	(4)	(5)	(6)
Paywall	-0.1718*** (0.0556)	-0.1878*** (0.0341)	-0.1121* (0.0586)	-0.2061** (0.0843)	-0.1543** (0.0551)	-0.1805*** (0.0484)
DMA-Newspaper FE	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓
Observations	168,457	78,336	216,095	213,414	158,193	127,485
R^2	0.95142	0.96119	0.94243	0.94366	0.95279	0.94574

Notes: Column (a) shows baseline estimates. Column (b) uses only newspaper-region units that are observed over all months. Columns (c) to (f) report individual (non-cumulative) changes to the baseline specification. Column (c) adds Newsday, Dallas Morning News, and Washington Post to the treatment group. Column (d) adds Wall Street Journal and New York Post to the control group. Column (e) excludes the Denver Post and Cleveland Star-Tribune. Column (f) uses data from Jan. 2008 to Dec. 2016. All regressions estimated using robust Sun and Abraham (2021) estimator. Standard errors clustered by newspaper. Significance levels: *: 0.1, **: 0.05, ***: 0.01.

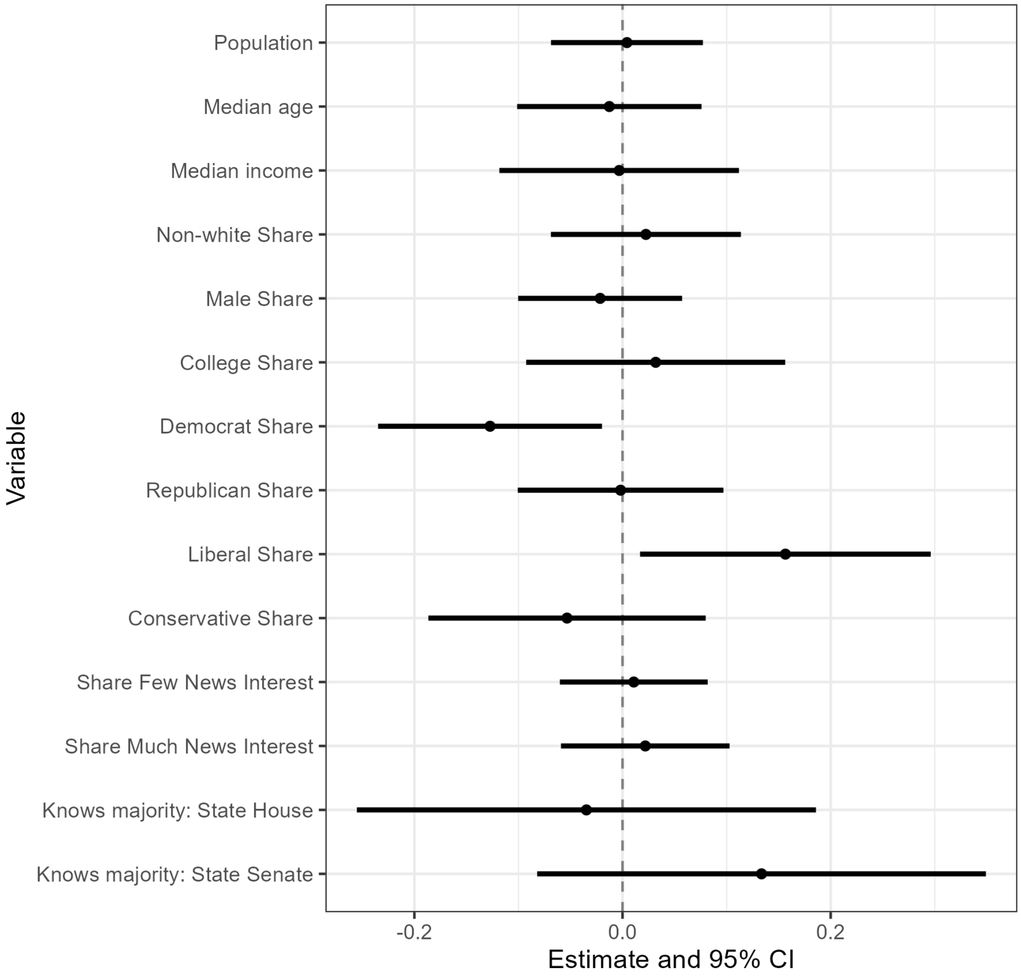
Appendix Table B2. Effect of paywalls on newspaper searches, by regional popularity of newspaper

Dependent Variable:	log(searches)			
Regional popularity:	Low		High	
Model:	(1)	(2)	(3)	(4)
Paywall	-0.2216*** (0.0541)	-0.1917*** (0.0638)	-0.1087* (0.0569)	-0.1242* (0.0612)
DMA-newspaper FE	✓	✓	✓	✓
Month FE	✓		✓	
DMA-month FE		✓		✓
Observations	140,949	140,949	131,538	131,538
R^2	0.95	0.96	0.95	0.96

Notes: In columns (1) and (2), observations in post-period are only DMAs with below-median pre-paywall popularity of the treated newspapers. Analogously with above-median popularity for columns (3) and (4). Standard errors clustered by newspaper. Significance levels: *: 0.1, **: 0.05, ***: 0.01.

C Appendix: Effect of paywalls on political knowledge

Appendix Figure C1. DMA-level correlates of paywall exposure in 2013



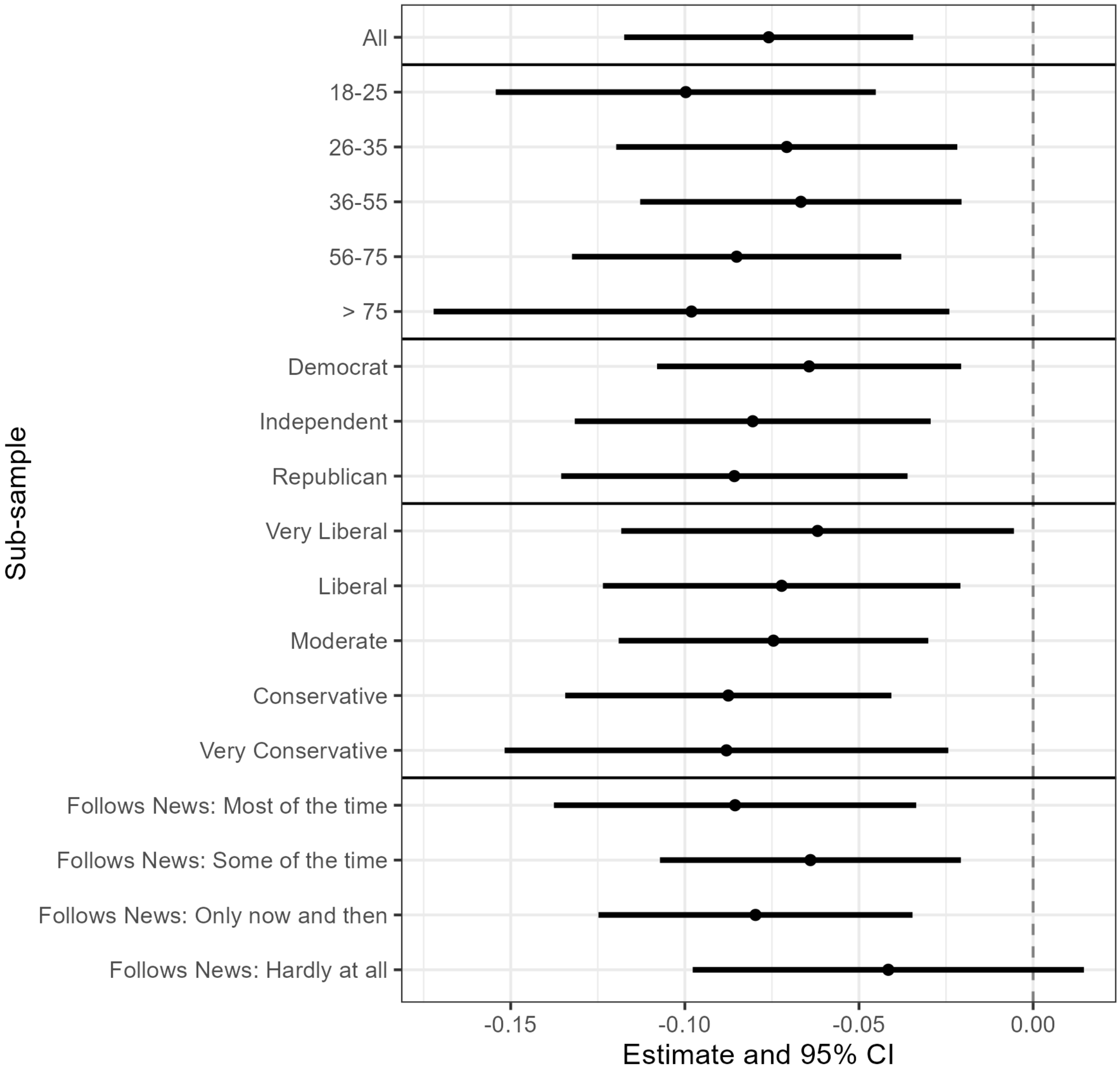
Notes: Point estimates and 95 percent confidence intervals from regression of (continuous) paywall exposure on DMA characteristics obtained from 2010 US Census (population, age, income, race, sex, college) and 2006–2010 CES (party identification, ideology, news interest, knowledge). Party identification derived from 3-point scale. Ideology derived from 5-point scale. News interest derived from 4-point scale (”Few”=1, ”Much”=4). All variables standardized. Robust standard errors.

Appendix Table C1. Balance in characteristics between regions with high and low paywall exposure, before 2011

2013 Paywall Exposure:	**Low**			**High**		
Variable	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**
Knows majority: State House	0.336	0.472	33,400	0.436	0.496	60,008
Knows majority: State Senate	0.369	0.482	33,431	0.457	0.498	60,056
Knows majority: US Senate	0.668	0.471	48,564	0.697	0.459	85,154
Knows majority: US House	0.692	0.462	49,622	0.719	0.449	86,922
Age	45.965	15.561	52,018	45.643	15.819	90,967
Female	0.531	0.499	52,018	0.506	0.500	90,967
White	0.768	0.422	52,018	0.721	0.449	90,967
Married	0.573	0.495	51,432	0.552	0.497	90,308
Has child	0.305	0.460	35,556	0.304	0.460	62,528
College	0.312	0.463	52,018	0.358	0.479	90,967
Family Income > 60k	0.334	0.472	52,018	0.383	0.486	90,967
Family Income > 100k	0.134	0.341	52,018	0.165	0.371	90,967
Employed	0.416	0.493	52,018	0.438	0.496	90,967
Unemployed	0.178	0.382	52,018	0.160	0.366	90,967
Retired	0.157	0.364	48,454	0.151	0.358	84,711
Democrat (pid3)	0.393	0.488	48,175	0.392	0.488	84,179
Independent (pid3)	0.301	0.459	48,175	0.310	0.462	84,179
Republican (pid3)	0.306	0.461	48,175	0.298	0.457	84,179
Liberal (ideo5)	0.212	0.408	52,018	0.240	0.427	90,967
Moderate (ideo5)	0.313	0.464	52,018	0.312	0.464	90,967
Conservative (ideo5)	0.338	0.473	52,018	0.323	0.468	90,967
Follows news: Hardly at all	0.053	0.224	52,018	0.050	0.218	90,967
Follows news: Sometimes	0.247	0.431	52,018	0.239	0.427	90,967
Follows news: Now and then	0.111	0.314	52,018	0.103	0.304	90,967
Follows news: Usually	0.368	0.482	52,018	0.387	0.487	90,967

Notes: Average mean, standard deviation, and number of observations for individual characteristics between 2006 through 2010 obtained from CES, separately for regions with below- and above-median paywall exposure in 2013. Party identification (pid3) derived from 3-point scale. Ideology (ideo5) derived from 5-point scale. News interest derived from 4-point scale.

Appendix Figure C2. Heterogeneous effects of paywall exposure (State Senate)



Notes: Point estimates and 95 percent confidence intervals for coefficient on $High\ PWX$ on knowledge of State House majorities in Equation (3) in different subsamples. The first section ("All") refers to the full set of observations. The second section refers to age. The third section refers to party identification. The fourth section refers to political ideology. The fifth section refers to the tendency of keeping up with the news. Survey weights are included. Standard errors clustered by DMA.