All the Headlines that Are Fit to Change*

Tin Cheuk Leung† Wake Forest University

Koleman Strumpf‡ Wake Forest University

November 2026

Abstract

How should a newspaper decide whether to revise a headline or learn which alternative performs better? We develop a Bayesian model in which editors retain the current headline, revise immediately, or conduct an A/B test comparing two headlines. Testing can improve the headline shown to later readers, but uses part of an article's finite audience before learning is complete and incurs resource costs. We construct high-frequency homepage histories for more than 56,000 New York Times articles, recording detected revisions and tests, subsequent headline choices, and reliably observed periods without a detected intervention. We link these histories to archived URL sharing and headline characteristics to estimate the static decision problem. The model explains overall testing and revision frequencies and much of the tendency to test when more readers are expected afterward. It is less successful in predicting which headline editors retain after a test and when interventions occur. Counterfactuals compare retention, revision alone, testing alone, and both options. Direct revision changes measured headline characteristics more than testing; final headlines can differ from those most readers encounter. Changes in negativity, partisan-language direction, and readability vary across news sections. A dynamic extension allows waiting for demand information or better testing opportunities; its structural estimation is reserved for a subsequent draft.

JEL Classification: D83, L29, L82
Keywords: A/B testing, media economics, Bayesian learning, attention, digital publishing

*We thank Luis Aguiar, Aimee Chin, John Dalton, Simone Galperti, Lisa George, Aeimit Lakdawala, Christian Peukert, David Rothschild, Andrea Szabo, and seminar participants at the CUHK–Zhejiang University Workshop on Media Economics, Penn Carey Economics of Digital Services Research Symposium, CREED, Microsoft Research Lab–New York City, NYC Media Seminar (Columbia and CUNY Hunter), SEA, University of Houston, Virtual Digital Economy seminar, and William and Mary for their comments. We thank Sean Anderson of the WFU HPC Facility, Tom Lam, and Julian TszKin Chan for help with coding, and Tom Cleveland for sharing his NYT headline data and scraper code. We gratefully acknowledge financial support from the Center for Technology, Innovation and Competition at the University of Pennsylvania and the Knight Foundation. The usual disclaimer applies.

†Department of Economics, Kirby Hall, Wake Forest University, Box 7505, Winston-Salem, NC 27109. Email: leungtc@wfu.edu.

‡Email: strumpks@wfu.edu.

"And now it's alright, it's okay
And you may look the other way
We can try to understand
The New York Times' effect on man"
from The Bee-Gees song, Stayin' Alive

1 Introduction

In March 2020, as Congress debated a coronavirus relief package, the New York Times (NYT) repeatedly changed a headline about the negotiations.[1] Wording emphasizing Democrats' decision to block the bill was revised to mention their demands for worker protections and then the partisan disagreement threatening a deal. President Trump publicly criticized the changes. The episode illustrates the flexibility of digital news: the same story can reach successive readers through different presentations. An observed headline change alone, however, does not reveal whether an editor acted on an existing judgment, compared alternatives experimentally, or waited for a better opportunity. With newspapers' shift to digital editions, such headline changes are now commonplace, with companies such as Chartbeat and Optimizely providing tools to create and assess different wording.

Why does a newspaper test alternative headlines, revise directly, or wait? Learning can lead to a headline that is more appealing to later readers, but at the cost of presenting initial readers with a suboptimal version while a comparison runs. Testing can also require resources to initiate. Costly firm learning, where immediate payoffs may be sacrificed for future gains, follows in the tradition of Rothschild's (1974) two-armed bandit model. An editor who expects an alternative title to greatly boost viewership may therefore switch immediately; if there is less certainty of a gain, the editor may choose to conduct an A/B test or wait to make a decision later. News makes this trade-off sensitive to timing. A busier experimental window supplies a larger sample, but postponement loses readers who could have benefited from earlier learning.

We develop a finite-audience Bayesian model, building on Azevedo et al. (2023) and Feit and Berman (2019), in which an editor has an incumbent headline and a possible alternative. The static choice is to retain the incumbent, switch directly, or conduct an A/B test. If the alternative is clearly favored, switching avoids the delay and divided audience of testing; if the

[1] See Cochrane, Tankersley, and Smialek (2020), March 22.

incumbent is favored, retention may be preferable. Testing is worthwhile when information could change the choice and the expected improvement for later readers covers its opportunity and resource costs. Readers arriving during the test supply information; readers arriving afterward benefit from it. Two equally informative experiments can therefore have different returns: learning is more valuable when many readers remain than when few do. Holding the alternative's expected performance, uncertainty, experimental capacity, and testing costs fixed, a larger post-test audience raises the return to experimentation. This motivates examining whether editors test more often when more readers are expected afterward, controlling for expected traffic during testing (Section 4.6). The dynamic extension adds waiting, preserving the option to reconsider. A busier future period may provide a more informative experiment, and new demand information may reveal that too few readers remain to warrant testing. Waiting also uses up readers who could benefit from earlier learning. We estimate the static decision problem here; dynamic estimation is reserved for a subsequent draft.

Our first empirical contribution is a record of editorial information acquisition. We reconstruct 61,546,862 high-frequency homepage observations for 56,714 NYT articles from July 28, 2021, to June 6, 2023. A persistent headline change indicates *direct revision*; *experimentation* can compare headlines that differ by only a few words and is detected through cycling before settlement. A fifth of articles have at least one detected headline change. We link these histories to archived timestamped URL sharing and measures of eight headline characteristics, including sentiment, partisan direction, readability, and specificity. A broad adoption sample describes which tested headlines editors retain. A separate first-event risk panel follows 36,234 articles through 1,380,928 eligible intervals, including periods without a detected intervention.

Our second empirical contribution is structural estimation of the finite-audience Bayesian model. An editor periodically reconsiders the current headline, choosing whether to retain it, revise it, or experiment. We do not observe this deliberation, but we observe actions, long stretches without a detected change, and the headline retained after completed tests. Direct revision is attractive when the editor already favors an alternative; experimentation is attractive when information could improve the choice for later readers. Adoption reveals an additional consequence of that learning. Fitting these observations jointly helps distinguish the cost of starting a test, prior expectations and uncertainty about alternatives, the frequency of reconsideration, and the distribution of available challengers. The model must explain

when editors act and what they select, as well as how often they test.

The model reproduces aggregate testing and revision frequencies reasonably well and generates much of the tendency to test when more readers are expected afterward. Matching the total number of tests alone would miss this economic implication: the same information is more valuable when it can improve choices for more readers. Adoption, early timing, and later-period testing remain less well matched. Dynamic estimation may help explain timing by measuring how editors balance better future opportunities against readers lost through delay.

Our third empirical contribution quantifies editorial decision-making in the era of digital newspapers. Holding the alternative set fixed, we compare headline retention, revision alone, experimentation alone, and both options on common article and audience paths. Direct revision has larger average effects on measured characteristics, while experimentation can improve the information used to select a headline. Experimenting may change final selection and average reader exposure in different directions because it delays settlement and displaces some immediate revisions. Effects on negativity, political-language direction, and readability also vary across U.S. news, World, and Opinion.

Section 3 develops the static and dynamic framework after the related literature. Section 4 describes measurement, adoption, forecasts, and the remaining-audience evidence. Section 5 estimates and evaluates the static model; Section 6 examines how testing and direct revision change headline selection, exposure-weighted presentation, and normalized editorial payoffs. Section 7 concludes.

2 Related Literature

The economics of A/B testing treats experimentation as an investment in a subsequent decision. Azevedo et al. (2023) derive implementation and sample-size rules with Gaussian priors and linear experimentation costs, showing why payoff-maximizing adoption can differ from conventional significance rules. Feit and Berman (2019) account explicitly for the finite population: people used to compare alternatives are no longer available for deployment of the selected treatment. Our framework applies these considerations to editorial learning, where both experimental traffic and the audience remaining afterward change rapidly. The histories reveal not only completed comparisons but also direct revisions and covered opportunities

without intervention.

Experimentation also affects organizational decisions and resource allocation. Athey and Luca (2019) describe the role of economics and experiments in technology companies, while Koning, Hasan, and Chatterji (2022) study experimentation and startup performance. News allows the timing of an individual information investment to be examined over hours rather than a product's much longer life. Separating the experimental audience from the deployment audience is particularly consequential when an article can lose much of its readership while information is acquired.

The timing problem connects to dynamic learning and optimal stopping. Keller and Rady (1999) study experimentation when the unknown state changes over time. Here, uncertainty about headline effectiveness is distinct from uncertainty about audience opportunities: testing reveals relative performance, while waiting can reveal demand or provide a more informative experiment. We develop that continuation problem but estimate immediate choices under repeated static reviews. A subsequent dynamic implementation would follow the discrete-choice approach of Rust (1987); the current estimator does not solve its Bellman problem.

Within news production, Hagar and Diakopoulos (2019) document through newsroom interviews how headline testing changes workflows and the use of audience metrics. Our histories complement this institutional evidence by distinguishing direct judgment from experimental information acquisition. Audience responses give those choices economic stakes. Randomized headline comparisons show that negative language can increase clicks (Robertson et al., 2023) and that readers favor simpler headlines (Shulman, Markowitz, and Rogers, 2024). Subtle headline misinformation can also affect interpretation of the article (Ecker et al., 2014). We consequently measure several headline characteristics jointly and examine the editor's final selection. Without private arm-level responses, these adoption associations do not identify causal click responses to individual attributes.

Media-slant research links presentation to reader demand and editorial incentives. Gentzkow and Shapiro (2006) model reputational incentives, and Gentzkow and Shapiro (2010) relate newspaper slant to demand and ownership. Leung and Strumpf (forthcoming) use homepage duration to study media bias in the NYT and WSJ. Shen (2026) examines how platform curation changes newspaper slant through editorial supply responses to algorithmic incentives. These studies connect the allocation of attention to editorial presentation. Our within-article histories instead reveal how editors acquire information and time their

choices among alternative headlines. The placement extension connects these margins when headline-generated engagement changes later exposure.

Within-article comparisons restrict the underlying story more closely than cross-outlet comparisons, but editorial motives and audience demand can still jointly shape wording. An alternative can be retained because it attracts readers, satisfies an editorial objective, or does both. Measuring political direction alongside other attributes places that possibility within a broader information-acquisition problem, without interpreting each politically associated phrase as evidence of a separate political motive.

Audience sorting gives the selection margin a broader role. Iyengar and Hahn (2009) show that perceived partisan affinity affects news choice, while Gentzkow and Shapiro (2011) compare ideological segregation across online news, offline news, and social interactions. Selective exposure is distinct from changes in attitudes; Prior (2013) reviews mixed evidence on partisan media and polarization. Headline experimentation supplies a channel through which outlets can learn about demand within a story and adjust its presentation. Whether learning broadens appeal or favors politically congenial audiences depends on whose responses are measured and how editors value them. We observe the selection stage, motivating counterfactual comparisons of selection and presentation while keeping changes in polarization outside the claims supported by our estimates.

3 Theoretical Framework

An editor can retain the current headline, switch to an alternative, or learn about their relative performance through a test. Learning can improve the headline shown to later readers, but uses part of a finite audience before the comparison is resolved. We combine Gaussian learning (Azevedo et al., 2023) with this audience constraint (Feit and Berman, 2019). The static model values immediate choices; the dynamic model adds the option to wait.

3.1 Headlines, information, and audience opportunities

An article i has an incumbent headline A and an available alternative, or challenger, B. The challenger may be private to the editor and persists within a content episode; information that changes the article can begin a new episode. Retaining A and switching immediately to

B have no incremental resource cost. Starting a test costs $k \geq 0$, excluding sunk platform fees and the audience opportunity cost derived below.

Each article has a finite flow of potential readers. Arrival rates vary with article age, time of day, story popularity, and news events, and future traffic need not be known exactly. Let r_{it} denote *baseline headline exposure opportunities* in interval t: chances to encounter the headline before accounting for its relative engagement payoff. The reference audience state $\mathcal{X}_{it}$ collects current demand conditions and the editor's information about them. Conditional on this state, the main model holds the audience process independent of the focal headline action, with innovations independent of private headline quality and experimental noise.[2]

Let T_i be the article's economic horizon and $\delta \in (0, 1]$ the discount factor per interval. Expected remaining discounted exposure is

$$F_{it} = \mathbb{E}_t \left[\sum_{j=0}^{T_i - t - 1} \delta^j r_{i,t+j} \right], \tag{1}$$

where $\mathbb{E}_t$ conditions on information available at t. For the deterministic, undiscounted comparison, total remaining audience is $N = n + M$: n exposures occur during the experiment and M afterward, so $F = N$. Exposure opportunities need not equal observed sharing or the newspaper's internal experimental assignments; Section 4 explains their empirical proxy.

The newspaper values engagement and editorial consequences. For headline $h \in \{A, B\}$, normalized payoff per opportunity is

$$u_i(h) = \vartheta_i' f(x_h, c_i) - \lambda_L L(h, c_i) + \xi_i \mathbf{1}\{h = B\}. \tag{2}$$

Here x_h contains headline characteristics, c_i is article context, and f allows their returns to depend on that context. The coefficient vector ϑ_i captures those returns; λ_L weights a possible loss L from inadequately supported presentation. The indicator $\mathbf{1}\{h = B\}$ equals one for the challenger and zero for the incumbent, so ξ_i is its unmeasured advantage relative to A. Characteristics include sentiment, partisan direction, information withholding, readability, specificity, personal consequences, agency, and certainty. Concrete detail, for example, may help a factual report but remove some of an essay's appeal.

For characteristic differences $d_i = f(x_B, c_i) - f(x_A, c_i)$ and loss difference $\Delta L_i = L(B, c_i) -$

[2]Appendix A.4 specializes this process to homepage display. Appendix H allows headline-generated engagement to change future exposure; Appendix H.1 discusses future estimation of that extension.

$L(A, c_i)$, the challenger's relative payoff is

$$\Delta_i = u_i(B) - u_i(A) = \vartheta_i' d_i - \lambda_L \Delta L_i + \xi_i, \tag{3}$$

$$\Delta_i \mid d_i, c_i \sim N(\mu_i, v_i), \quad \mu_i = \beta' d_i - \lambda_L \Delta L_i, \quad v_i = d_i' \Sigma d_i + \sigma_\xi^2. \tag{4}$$

Conditional on d_i, c_i, the coefficients have mean β and covariance Σ; ξ_i has mean zero and variance σ_ξ^2 and is uncorrelated with ϑ_i. The prior mean μ_i is the editor's expected challenger advantage, while the prior variance $v_i > 0$ measures uncertainty about it. Relative headline payoffs are held fixed as audience conditions change. An editor who strongly favors B may switch directly, while one who favors A may retain it. Near an uncertain ranking, information has more scope to change the decision. Large characteristic differences can alter both expected advantage and uncertainty, so they need not increase testing.

3.2 What an experiment teaches the editor

Suppress article and time subscripts when unnecessary. A balanced experiment assigns half of its $n > 0$ exposure units to each headline and produces a noisy comparison Z of their relative payoff:

$$Z = \Delta + \epsilon, \qquad \epsilon \sim N(0, \nu^2/n), \qquad \Delta \sim N(\mu, v). \tag{5}$$

The parameter $\nu^2 > 0$ governs signal noise, which is independent of the prior conditional on the state. More experimental readers make the comparison less noisy. When the experiment measures engagement, known editorial-loss differences must be subtracted to obtain a signal of net advantage; uncertain losses require additional beliefs.

After observing realization z, the posterior mean $m(z)$ and posterior variance v^+ are

$$m(z) = \mu + \frac{v}{v + \nu^2/n}(z - \mu), \tag{6}$$

$$v^+ = (v^{-1} + n/\nu^2)^{-1}. \tag{7}$$

A more precise signal receives greater weight. Before observing it,

$$m(Z) \sim N(\mu, s^2), \qquad s^2 = \frac{v^2}{v + \nu^2/n} = v - v^+. \tag{8}$$

Thus s^2 measures uncertainty resolved by testing. A precise experiment leaves little uncertainty v^+ afterward, but its possible results can produce widely differing posterior means. That dispersion is valuable because the editor can respond to the evidence. An uninformative experiment leaves the prior decision almost unchanged.

With zero switching cost and a common future audience, the editor adopts B when $m(z) > 0$. Relative to always displaying A, optimal post-test payoff per reader is

$$J(\mu, s) = \mathbb{E}[\max\{0, m(Z)\}] = \mu\Phi(\mu/s) + s\phi(\mu/s), \tag{9}$$

where ϕ and Φ are the standard normal density and cumulative distribution function. The editor acts on favorable evidence and declines B after unfavorable evidence. Compared with the best immediate choice, $\mu^+ = \max\{0, \mu\}$, information is worth

$$G(\mu, s) = J(\mu, s) - \mu^+ = s\phi(|\mu|/s) - |\mu|\Phi(-|\mu|/s) \geq 0. \tag{10}$$

Set $J = \mu^+$ and $G = 0$ at $s = 0$. For $s > 0$,

$$\frac{\partial G}{\partial s} = \phi(|\mu|/s) > 0, \qquad \frac{\partial G}{\partial |\mu|} = -\Phi(-|\mu|/s) < 0. \tag{11}$$

Information is most valuable near prior indifference (Figure 1). If the editor already strongly favors one headline, a test is unlikely to reverse its choice.

Adoption maximizes expected payoff rather than statistical significance. Before testing, its probability is $\Phi(\mu/s)$; after a result, the probability that B is truly better is $\Phi(m(z)/\sqrt{v^+})$. A risk-neutral editor can adopt a favorable posterior mean without requiring a conventional significance threshold. The start cost is sunk by adoption, although it selects which pairs are tested. Appendix A gives the derivations.

3.3 The static choice

Let Q denote an action's expected incremental payoff relative to retaining A for all remaining readers. Subscripts K, S, and T denote keeping, switching, and testing. Keeping earns zero relative to this benchmark. Switching gives the challenger to every remaining reader, earning its expected advantage over the whole audience. Testing gives half the experimental audience to each headline, then uses the noisy comparison to select the headline with the

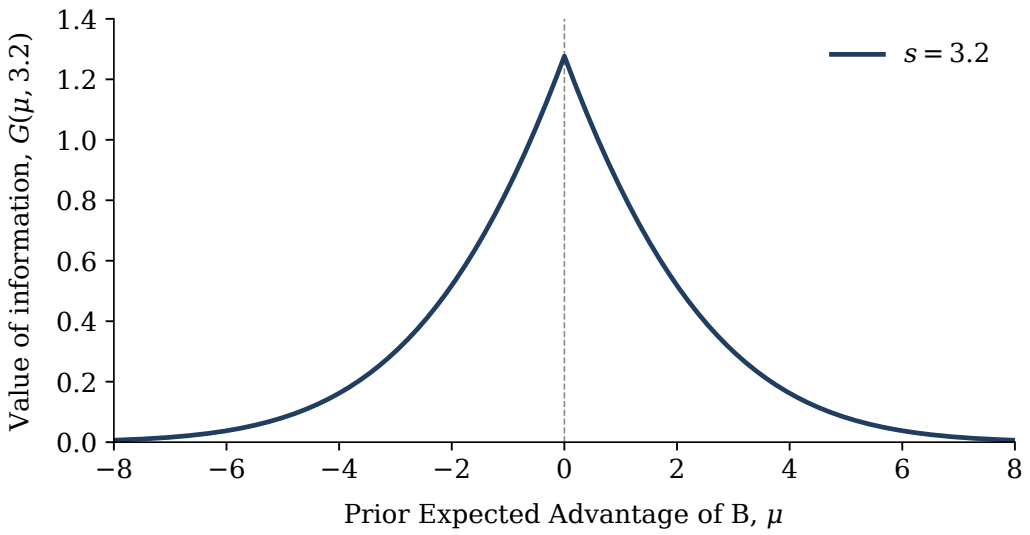


Figure 1: Information is most valuable near prior indifference

Theoretical value $G(\mu, 3.2)$ per later exposure relative to the best immediate choice. The hypothetical $s = 3.2$ is the standard deviation of possible posterior means. The curve excludes experimental allocation and resource costs and is not estimated from NYT data.

higher posterior expected payoff for later readers, net of the start cost. If N is known, the balanced test completes before readership ends, and $\delta = 1$,

$$Q_K = 0, \tag{12}$$

$$Q_S = N\mu, \tag{13}$$

$$Q_T(n; k) = \frac{n\mu}{2} + MJ(\mu, s(n)) - k. \tag{14}$$

Proposition 1 (Testing with a finite audience)**.** *An informative feasible test strictly dominates the best immediate deployment decision if and only if*

$$MG(\mu, s(n)) - \frac{n|\mu|}{2} - k > 0. \tag{15}$$

Holding n, the prior, and signal precision fixed, its advantage rises with M. When $G > 0$, the required post-test audience is

$$M > M^*(\mu, n; k) = \frac{k + n|\mu|/2}{G(\mu, s(n))}. \tag{16}$$

Subtracting $N\mu^+$ from the test value proves the result. The first term is the gain from better choices for later readers. The second is the loss from allocating half the experimental

audience to the initially disfavored headline. When $\mu > 0$, testing must improve on switching immediately; when $\mu < 0$, it must improve on keeping A. The resource cost k is a separate input, not another charge for those same readers.

At prior indifference, testing requires $Ms/\sqrt{2\pi} > k$. If $k = 0$, any informative feasible test with $M > 0$ improves on immediate commitment at a tie. A positive cost introduces a minimum scale even there. More experimental traffic improves precision but also increases allocation losses and, at fixed N, leaves fewer readers afterward. Appendix A derives the sample-size condition. The next subsection allows a fixed-duration test to receive an uncertain sample.

3.4 The dynamic model: waiting and the timing of intervention

The editor can also retain A now and reconsider later. The decision state is $S_t = (\mathcal{X}_t, d_i, \mu_t, v_t)$, combining reference audience conditions, the available candidate, and its prior. The candidate and prior persist while waiting; new demand information updates $\mathcal{X}_t$, not relative quality against an unseen challenger. Switching or test completion commits to the selected headline for the rest of the episode.

In practice, a test may last a fixed $L_T > 0$ review intervals rather than use a fixed sample. A busy interval yields more information than a quiet one. Along audience path Γ, let n_T count experimental exposures and D_T weight those same exposures for discounted payoffs:

$$D_T(\Gamma) = \sum_{j=0}^{L_T-1} \delta^j r_{t+j}, \qquad n_T(\Gamma) = \sum_{j=0}^{L_T-1} r_{t+j}. \tag{17}$$

These describe the same readers: n_T determines precision, while D_T discounts their payoffs. They coincide without discounting. Immediate switching and testing have values

$$Q_S(S_t) = \mu_t F_t, \tag{18}$$

$$Q_T(S_t; k) = -k + \frac{\mu_t}{2}\mathbb{E}[D_T(\Gamma) \mid S_t] + \delta^{L_T}\mathbb{E}[F_{t+L_T} J\{\mu_t, s(n_T(\Gamma))\} \mid S_t]\,. \tag{19}$$

The joint expectation values each traffic path's experimental information and remaining audience before averaging. These audiences can move together as demand rises or fades.

Evaluating information at the average sample and multiplying by average later exposure generally misses both that dependence and nonlinear information value.

Let V_t be optimal remaining incremental payoff. The subscript W denotes waiting: it earns zero current incremental payoff while preserving future choices, unlike static retention K. With a reference state sufficient to predict audience transitions,

$$Q_W(S_t; k) = \delta \, \mathbb{E}[V_{t+1}(\mathcal{X}_{t+1}, d_i, \mu_t, v_t; k) \mid \mathcal{X}_t], \tag{20}$$

$$V_t(S_t; k) = \max\{Q_W(S_t; k), Q_S(S_t), Q_T(S_t; k)\}. \tag{21}$$

Here $V_{t+1}(\mathcal{X}_{t+1}, d_i, \mu_t, v_t; k)$ evaluates the next decision state with the candidate and prior unchanged. Remaining value is zero at the economic horizon.

Waiting has competing effects. A busier future testing window can improve precision, and new demand information can avoid spending k on a test with little subsequent audience. Delay also loses readers who could benefit from earlier selection. An anticipated surge may therefore favor testing now, so the preferred headline is ready when those readers arrive. The Bellman comparison weighs these possibilities jointly.

If the same fixed-sample experiment remains feasible, beliefs and precision are fixed, no information arrives while waiting, and remaining audience only shrinks, testing is weakly better earlier. A constant resource cost alone cannot explain delay. Appendix A proves this limiting result and gives the homepage-specific transition representation. A recurrent extension would preserve further intervention options and posterior beliefs after an early test.

3.5 Implications for empirical analysis and estimation

The model separates readers who generate information from those who benefit afterward. Holding prior beliefs, experimental information, and resource cost fixed, equation (15) makes a larger later audience more valuable for testing. More experimental traffic also increases allocation costs. Section 4.6 therefore examines testing and expected remaining attention while controlling for experimental capacity. Forecast URL sharing proxies audience opportunities; actual headline exposures are unobserved.

Section 5 estimates the static choices jointly from testing, direct revision, eligible intervals without detected intervention, and headline selection after completed tests. Their joint patterns help distinguish candidate attractiveness, prior uncertainty, review frequency, and

testing cost. Observability restricts the empirical histories, as described there. The fitted model then generates Section 6's comparisons of available actions, final headlines, reader exposure, and normalized editorial payoff. Dynamic structural estimation, including audience transitions and the continuation value of waiting, remains for a subsequent draft.

4 Data and Measurement

High-frequency NYT homepage histories identify displayed headlines and editorial events. Linked URL-sharing records measure attention before and after those events. Article metadata, headline attributes, and related-topic activity connect the sources into a decision-risk panel.

4.1 Homepage histories and editorial events

We developed our homepage-history dataset by operating a web scraper at approximately one-minute intervals, using scraper code and headline data shared by Tom Cleveland. Each visit records the displayed headline and stable article identifiers. Associated metadata supply URLs, publication time, section, tone, and article information. Randomized user agents and omission of cookie headers reduce persistent assignment to one experimental arm. The resulting archive contains 61,546,862 records for 56,714 articles from July 28, 2021, to June 6, 2023. It also underlies the homepage-duration analysis in Leung and Strumpf (forthcoming); here we reconstruct headline histories and identify editorial events. We distinguish text states before scoring, so two titles with identical attribute scores remain different alternatives.

We infer editorial events from the sequence of displayed headlines. A change from A to B that remains at B is classified as a direct revision. Alternation between A and B before one headline is retained identifies an A/B episode; cycling among more headlines identifies multi-headline experimentation. Baseline classification requires thirty minutes of subsequent stability, and completed A–B–A reversions count as incumbent victories. We observe displayed headlines rather than the newspaper's own records of test starts and completions. Changes cannot be followed while an article is absent or scraping fails; gaps exceeding five minutes break the observed spell. A returning article reveals its current headline but not necessarily the changes made during its absence.

Table 1 separates episodes from affected articles. Direct revisions involve 15.57 percent of archive articles; two-way and multi-headline tests together involve 7.09 percent. At least

one classified intervention affects 20.31 percent, and 6.42 percent have repeated interventions. Categories overlap across articles. These detected frequencies cover the classifier's first seventy-two hours rather than lifetime editing.

Table 1: Detected interventions: episodes and distinct articles

Event	Episodes	Articles	Articles (%)	Episodes/article
Direct revision	11,309	8,831	15.57	0.199
Two-way test	5,279	3,676	6.48	0.093
Multi-headline test	575	540	0.95	0.010
Any test	5,854	4,019	7.09	0.103
Any classified intervention	17,163	11,518	20.31	0.303

Percentages and episodes/article use all 56,714 archive articles; 53,316 have histories within the classifier's post-publication window. Any test includes two-way and multi-headline episodes. Categories overlap. Unresolved and ambiguous sequences are excluded; no detected event does not establish no hidden action.

Figure 2 shows publication age at detected intervention. A displayed change occurs between the last observation of the incumbent and the first observation of a new headline; we assign its time to the midpoint. If A is observed at 9:01 and B at 9:02, the assigned time is 9:01:30. The median interval is approximately one minute, and all intervals used here are shorter than five minutes. This dates the first observed change, not settlement or the newspaper's unobserved internal start of randomization. Tests and revisions occur beyond the initial publication period: 57.76 and 50.14 percent, respectively, occur at least six hours later; 10.17 and 6.66 percent occur after twenty-four hours. Median ages are 7.45 and 6.02 hours. Revisions are more front-loaded, although first-classified-intervention medians are similar at 4.91 and 4.83 hours.

The timing distribution reflects detection, homepage entry, and repeated editing as well as editorial incentives; it is not causal evidence for waiting. Later tests nevertheless retain an audience over which learning can pay off. The hazard analysis below applies additional risk restrictions, while adoption requires observed completion. All 484 incomplete two-/multi-way episodes retain missing winners, including 348 ending at a spell boundary. Appendix B.1 documents these sample distinctions and the construction rules.

4.2 URL-sharing records

We search article URLs through Twitter's academic API. The processed archive contains 26,623,103 article–tweet records for 82,367 articles, a wider universe than the homepage

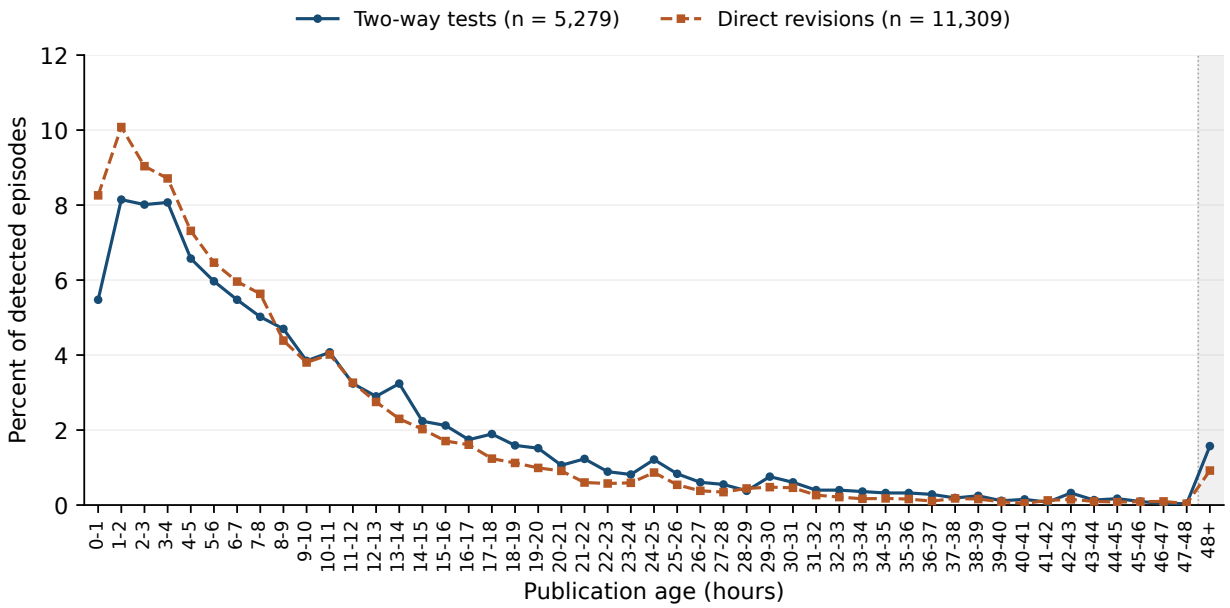


Figure 2: Publication age at detected intervention

Shares of 5,279 two-way tests and 11,309 revisions, including repeated interventions, in hourly bins from 0–1 through 47–48 hours, followed by 48+ (pooling 48–72 hours). Bins are left-closed and right-open. Age uses the midpoint between the last incumbent observation and the first observation of the changed headline. These are event distributions, not hazards conditional on eligibility.

sample. Timestamps, author identifiers, and sharing-type fields permit counts of original shares, retweets, and distinct accounts. Strictly earlier shares measure current attention and forecast future attention; subsequent records supply outcomes, including after homepage departure. Sharing measures attention rather than clicks, readership, or experimental impressions. Acquisition-success logs are unavailable, so zeros mean no archived record rather than certified complete collection.

Figure 3 follows all 56,714 homepage-linked articles with valid publication times for seventy-two hours, retaining zero-share hours. Mean sharing falls from 32.62 in the first hour to 4.98 in hours 24–25 and 0.71 in hours 71–72. Clock-time rates use article-hours as their denominator. To separate daily timing from article-age composition, we also average each clock hour's age-specific rates over a common, equally weighted distribution of the seventy-two age hours. The adjusted profile peaks around 10 a.m. and is lower overnight; the daily pattern persists after age adjustment.

The decay limits the audience that can benefit from learning. Daily variation also means equally long tests can provide different samples. Testing during a busy window yields more information, but waiting for that window uses up readers who might otherwise benefit from

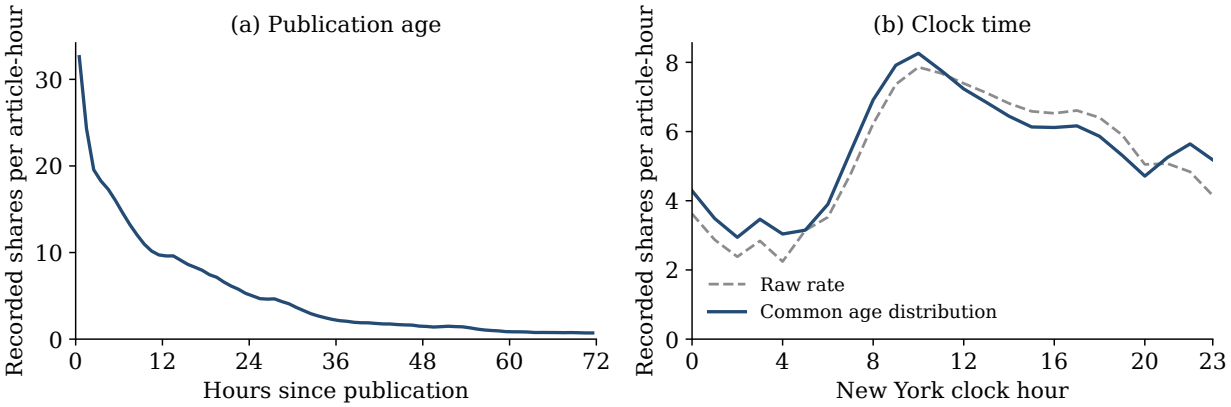


Figure 3: Article attention by publication age and New York clock time

21,605,755 recorded shares over 4,083,408 article-hours, including 2,367,347 zero-share hours. Left: hourly publication-age means. Right: shares per article-hour at actual New York clock time, accounting for daylight saving; the adjusted line uses common age weights. Fractional hours at clock boundaries enter the denominator. Zeros denote no archived record. Appendix B.2 gives construction details.

earlier selection. These patterns motivate separate experimental and later audiences and the dynamic option to wait.

4.3 Headline attributes

We measure eight headline characteristics. Sentiment is the positive-minus-negative probability from a pinned Twitter-RoBERTa model on $[-1, 1]$. Partisan direction measures language associated with Democratic rather than Republican politicians. We link ProPublica party affiliations to 2,100 politician accounts, yielding 4.14 million party-linked raw tweet records. After filtering institutional accounts, duplicates, and conflicting texts, 275,770 tweets remain for classifier training and evaluation. Words, word combinations, and short character sequences describe each text; supervised learning finds patterns associated with each party. Calibration turns the classifier output into a party-language probability. Evaluation holds out entire politicians, linking campaign and official accounts so the same person's language cannot appear in both training and evaluation. Accuracy is 72.34 percent on held-out politicians and 65.07 percent on titles shared by unseen politicians, supporting transfer to headline text.

Applied to a headline, a score near one indicates Democratic-associated language and a score near zero Republican-associated language. This is language direction, rather than newspaper bias or headline extremity. Names and events can affect political interpretation. The adoption analysis retains its established context-assisted measure, which assigns neutrality

where supported U.S.-political context is absent. That context snapshot is not certified as available at the editorial decision time; initiation and structural analyses instead use the headline-only counterpart. Appendix B.3 details filtering, calibration, and validation.

Six additional attributes describe the information offered by a headline. Information withholding measures an unresolved gap that can invite reading (Loewenstein, 1994). Readability captures ease of comprehension, motivated by evidence that simpler headlines attract engagement (Shulman, Markowitz, and Rogers, 2024). Specificity measures concrete detail, whose relationship with clicks can depend on baseline concreteness (Aubin Le Quéré and Matias, 2025). Personal consequences capture tangible implications for people's lives, connecting to the role of personal relevance in processing information (Petty, Cacioppo, and Goldman, 1981). Agency identifies responsibility for an action (Iyengar, 1991); certainty measures categorical assertion rather than qualification, which can also affect trust (Jensen, 2008). These dimensions can have different returns across contexts and do not rank overall headline quality.

A pinned local language model assigns probabilities to five fixed rubric categories for each dimension. Their expected category divided by four yields a score on $[0, 1]$, with higher values indicating more of the named attribute. The scale uses no sample percentiles. Context-assisted withholding can use a historically recorded excerpt subject to textual-entailment screening; the other five dimensions use wording alone. Adoption retains contextual withholding where available; initiation and structural inputs use its headline-only counterpart. Appendix B.3 gives the rubrics.

Challenger–incumbent score differences are small on average but dispersed. In completed pairs, 41.3 percent differ by more than 0.1 in readability and 41.2 percent in specificity; only 11.7 percent do so in partisan direction. These distributions matter when interpreting effects per 0.1 unit, particularly in sections with little partisan variation.

4.3.1 Headline attributes and selection after testing

For completed two-way episode r, we relate challenger adoption to joint attribute differences:

$$\Pr(B \text{ adopted} \mid r) = \Phi\left[\sum_{\ell=1}^{8} b_\ell (f_{Br\ell} - f_{Ar\ell}) + W_r'\gamma\right]. \tag{22}$$

Controls W_r are news/opinion indicators, an initial political-keyword indicator, log article age, and year indicators; inference clusters by article. The sample has 4,805 episodes from 3,453 articles, with challengers retained in 59.56 percent. It conditions on observed completion among selected candidate pairs, rather than identifying demand for arbitrary headlines. The private arm-level results are unobserved.

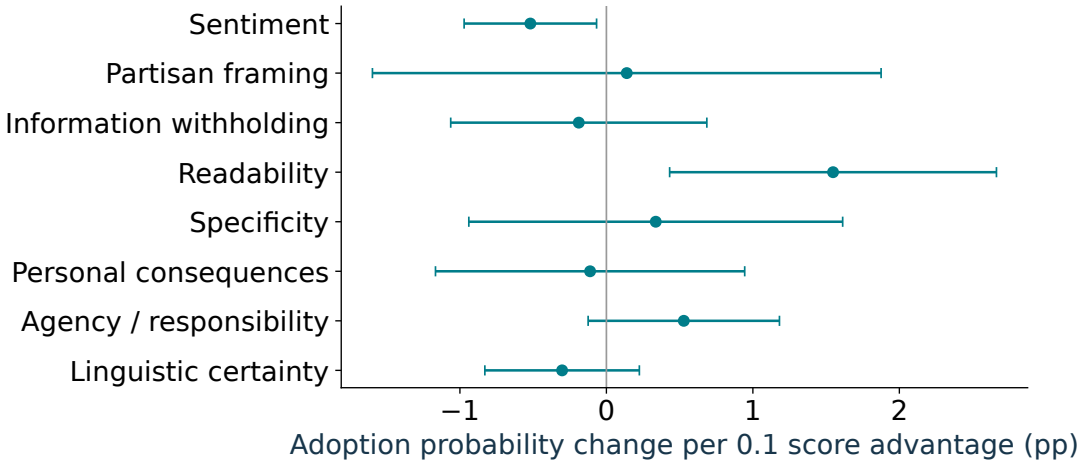


Figure 4: Headline attributes and challenger adoption

Pooled probit average marginal effects in percentage points per 0.1-unit challenger–incumbent difference, with article-clustered 95 percent intervals. All eight attributes and equation (22) controls enter jointly. "Curiosity" is information withholding; "personal relevance" is personal consequences.

Readability has the clearest pooled association (Figure 4). A 0.1-unit advantage raises estimated adoption by 1.55 percentage points [0.43, 2.66], about 2.6 percent of the observed adoption rate. This is a moderate selection shift at a difference common in the sample; the local effect need not scale linearly to larger changes. More positive sentiment is associated with 0.52 points lower adoption [−0.97, −0.07]. The six additional attributes are jointly informative ($p = 0.012$), while most individual intervals include zero. Relatively negative wording is thus more likely to be selected among tested pairs, alongside an advantage for clarity. Neither result identifies the causal effect of an attribute on engagement.

Figure 5 shows how selection varies across sections. Specificity is favored in U.S. news and disfavored in Opinion; readability is positive in several sections but negative in Sports. Partisan-direction estimates differ from the pooled readability result. Per 0.1-unit shift toward Democratic-associated language, the full-sample adoption association is −2.75 percentage points in U.S. news $[-5.47, -0.04]$, 10.11 in World $[-0.07, 20.29]$, and 1.91 in Opinion $[-0.58, 4.40]$. Only the U.S. interval excludes zero nominally; none survives adjustment across

the 64 section–attribute effects. Among first interventions, the respective estimates shrink to −1.39, 4.75, and 1.08 points, with all three intervals including zero (Appendix C).

These differences can reflect the alternatives considered, the stories they describe, and how political language works in each context. A party-linked phrase may clarify domestic responsibility, identify an actor in international news, or express an argument in Opinion. The regressions do not identify these explanations. They characterize contextual selection among tested pairs, whereas the pooled readability advantage is a broader association with ease of comprehension. Neither analysis establishes universal reader preferences.

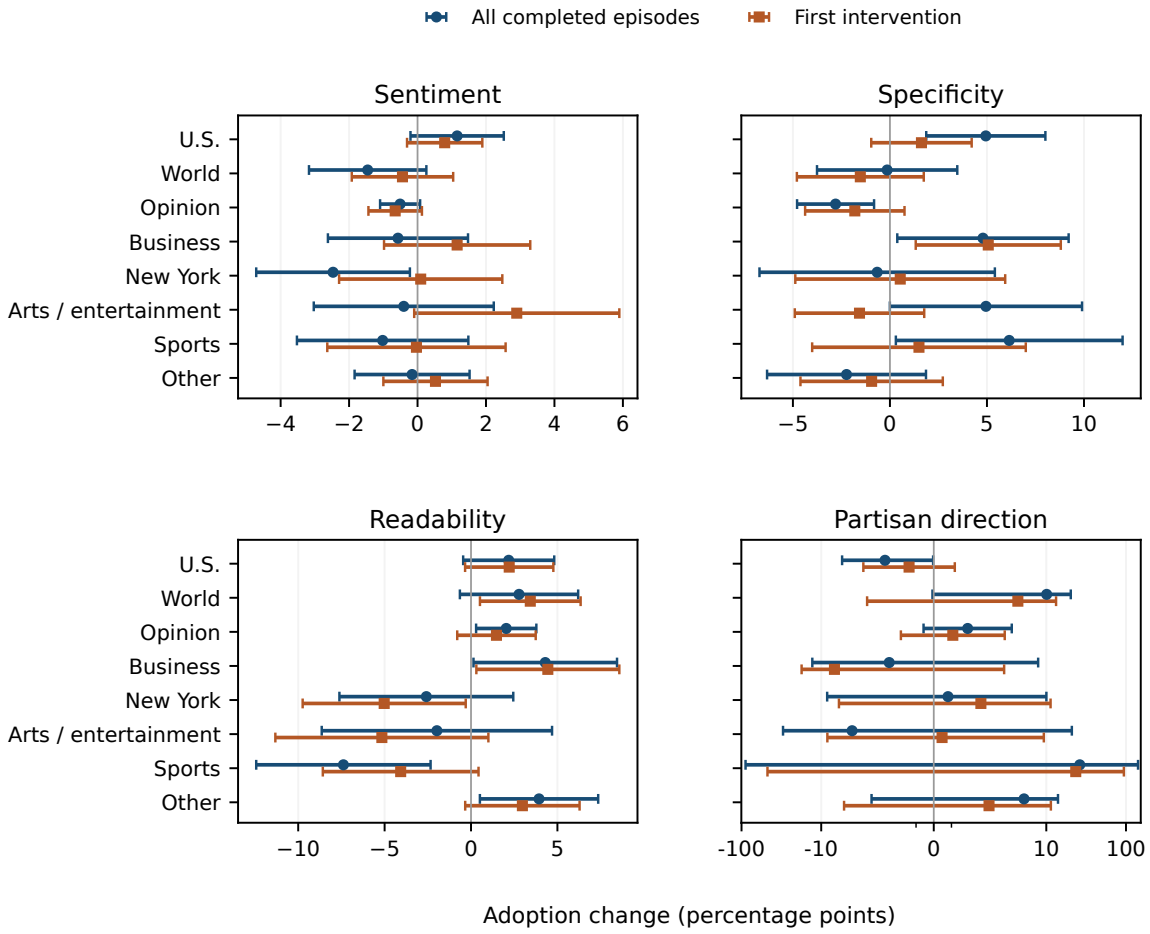


Figure 5: Headline selection across news sections

Four attributes in a two-by-two layout. Average marginal effects per 0.1 score unit with article-clustered 95 percent intervals: 4,805 full-sample episodes and 2,692 first interventions. All eight attributes are section-interacted. Sentiment is signed: higher means more positive. The partisan axis is linear within ±5 points and logarithmic outside; positive values favor Democratic-associated language.

The broad adoption population differs from the structural sample. Section 5 uses first detected binary cycling within eligible risk histories and headline-only score inputs. It does

not pool the 4,805 completed episodes into its initiation likelihood. Private results and priors are unobserved in both analyses, and differences in eligibility, detection, and score context prevent a coefficient-by-coefficient comparison.

4.4 The decision-risk panel

To study when editors initiate testing or revision, we need to observe both interventions and periods without a detected change. Articles that have already completed an experiment cannot by themselves reveal how initiation varies with expected future attention. We therefore follow articles while they remain eligible for a first detected intervention and their headlines are reliably observed on the homepage. Each observation is an article's fifteen-minute decision interval, recording a qualifying test, a direct revision, or no detected intervention, together with information available before the interval begins.

We start from the first observed homepage spell and require at least ten minutes of observed spell age and a stable incumbent headline, adequate scrape coverage, and no active detected experiment. A usable incumbent headline and the other clean-observation conditions establish which headline is currently in use so that a subsequent change can be recognized. The ten-minute requirement concerns observed homepage entry and incumbent stability, not publication age; publication ten minutes earlier is insufficient. Eligibility also requires available forecasts and publication age no greater than forty-eight hours.

For example, an article first observed at 9:00 with an unchanged headline can begin contributing decision intervals after 9:10, subject to coverage and the remaining conditions. Entry occurs at an eligible grid point, rather than automatically at 9:10. If a test is first detected at 10:05, earlier eligible intervals record no detected intervention. The final interval records the test and ends at the estimated event time; the article then exits the baseline panel. Truncating that interval avoids assigning time after the event to the initiation decision.

Successive intervals stop at the first qualifying intervention, a relevant ambiguity or content-change boundary, loss of reliable coverage, departure from the eligible spell, or the administrative boundary. Weakly classified first tests censor the strong-evidence history. We stop at the first intervention because testing or revision can change the incumbent and the editor's information; later decisions require tracking that evolving state. This is the first detected intervention in the defined observed history, not necessarily the article's first lifetime editorial action. Departure is a censoring boundary: editing may continue, but off-homepage

periods supply no observations of inaction.

The baseline contains 1,380,928 positive-exposure intervals from 36,234 articles, including 964 strong-evidence test initiations and 4,613 direct revisions.[3] All eligible non-event intervals are retained with weight one. Informative regression counts differ because dates with no event of the relevant type contribute no slope information to the date-intercept likelihood; Appendix Table A1 traces the sample. Linking each interval to current sharing and audience forecasts lets us examine initiation at different audience opportunities and estimate the intervention history jointly with adoption.

4.5 Forecasting the attention available before a decision

The value of testing depends on the audience that can participate and the audience that remains to benefit from a better headline. Neither has arrived when the editor acts. We therefore construct forecasts from predecision information, rather than using realized future shares as though the editor knew them in advance.

For each eligible decision, we predict sharing during the next thirty minutes and from thirty minutes afterward through twenty-four hours after the decision. The first proxies experimental capacity: more participating readers can produce a less noisy comparison of the headlines. The second proxies the later audience that can benefit from selection. Thirty minutes approximates an experimental window; it does not measure the NYT's actual randomized sample. For decision time t and predecision information $\mathcal{I}_{it}^{-}$, these forecasts are

$$\widehat{n}_{it} = \widehat{\mathbb{E}}[Y_i[t, t+0.5h) \mid \mathcal{I}_{it}^{-}], \qquad \widehat{M}_{it} = \widehat{\mathbb{E}}[Y_i[t+0.5h, t+24h) \mid \mathcal{I}_{it}^{-}]. \tag{23}$$

Here $Y_i[a, b)$ counts recorded URL shares in the indicated interval. Both targets include sharing after homepage departure. The forty-eight-hour decision-age ceiling keeps the common $t + 24h$ endpoint within the seventy-two-hour archive. A first-spell display-count sensitivity is bounded by total sharing and excludes later returns; it is not all future on-homepage exposure.

Predictions use article age, time of day and day of week, section and other article

[3]Of the 56,714 homepage articles, the observation and entry rules leave 46,790 with eligible risk intervals. The largest single reduction comes from chronological forecast availability: requiring forecasts trained only on earlier data excludes the initial training period and leaves 37,676 articles. Restricting decision times to publication ages no greater than forty-eight hours yields 36,234.

characteristics, strictly earlier own sharing, and historical or current related-topic activity. These variables describe both the article's stage in the news cycle and current interest in its subject. Two articles with equal cumulative shares can face different future audiences if one is attracting new attention while the other's sharing occurred much earlier. Recent sharing rates and momentum help distinguish those cases. Counts precede the prediction time, and rate denominators use only the history available since publication. Appendix D records the metadata-availability assumptions and implementation details.

We estimate the forecasting relationships on historical articles and evaluate them on later periods not used for fitting. Five expanding training vintages preserve this chronological separation. A baseline forecast uses article characteristics, cumulative popularity, and recent own sharing. An enhanced forecast adds more detailed recent rates and momentum. The preferred policy chooses between them within article-age groups using only earlier validation outcomes whose follow-up windows have ended; the first vintage uses the baseline. Editorial outcomes do not select the policy. Nonnegative predictions for successive quarter-hours add to the longer windows, keeping experimental and post-test audiences consistent. Nonoverlapping windows preserve a common twenty-four-hour forecast endpoint.

Validation compares predicted shares with recorded subsequent shares using Poisson deviance, a measure of count-prediction error for which lower values are better. The preferred twenty-four-hour loss is 211.09 versus 213.65 for the baseline, a 1.2 percent reduction; the one-hour reduction is 4.1 percent and is more precisely estimated. Own-article sharing history contributes much more predictive information than the modest additional gain from detailed rates and momentum. Appendix Table A5 reports the chronological comparisons.

The forecasts supply deterministic expected audience opportunities to the static estimator. Inserting conditional means into the learning value is a maintained approximation to uncertain future traffic, not integration over the full joint stochastic law in equation (19). The dynamic framework allows that richer process; we do not structurally estimate it here. Forecasts describe attention under historical editorial behavior, so their counterfactual use requires the reference-path assumptions in Section 6.

4.6 Remaining attention and test initiation

At a given experimental capacity, a larger later audience raises the return to the same information under the static model. We examine this implication using a piecewise-exponential

testing hazard:

$$\mathbb{E}[D_{it}^{T} \mid X_{it}] = e_{it} \exp\{\alpha_{d(t)} + g(\widehat{n}_{it}, \widehat{M}_{it}) + X_{it}'b\}. \tag{24}$$

The offset e_{it} is actual eligible time in hours. Date intercepts absorb daily conditions. Controls include current capacity, earlier sharing and momentum, age and spell conditions, topic activity, coverage, clock/calendar states, article characteristics, and incumbent scores. The audience function uses a spline in log-one-plus remaining attention, quadratic log capacity, and their interaction. Appendix D.1 records the full specification.

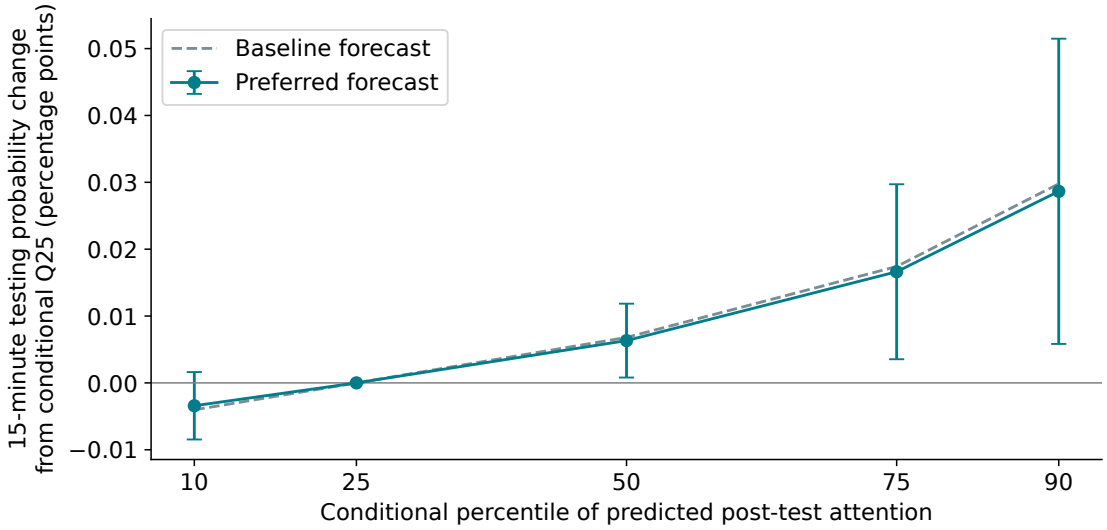


Figure 6: Testing and expected remaining attention

Original strong-evidence all-test definition. Adjusted fifteen-minute probability differences relative to conditional Q25, in percentage points, with article-clustered 95 percent intervals. Preferred forecasts select between baseline and enhanced forecasts using earlier chronological validation. The likelihood has 943,800 informative intervals from 27,658 articles; 943,359 support the preferred contrast.

Moving remaining attention from conditional Q25 to Q75 raises fitted fifteen-minute testing probability from 0.0953 to 0.1119 percent, a 0.0166-percentage-point change or 17.4 percent relative increase. Conditional cells share capacity decile, age band, and topic and require thirty observations. The article-clustered interval is [0.0035, 0.0297] points; article/calendar-week inference gives [0.0030, 0.0302]. These are net cause-specific interval probabilities, not article-lifetime experimentation rates.

Separating capacity matters: popular articles could be tested because they provide more information during the test, because more readers can use it later, or both. This comparison isolates observed variation in later stakes, while private priors and challengers remain uncontrolled. The association is consistent with the value-of-information mechanism,

rather than identifying a causal audience response or a testing-cost parameter. Section 5 evaluates whether the fitted model reproduces it under matched event definitions. Adding weak binary cycles changes the empirical all-test contrast to 0.01543 points; restricting that expanded definition to binary tests gives 0.01052. We retain these distinct benchmarks rather than compare incompatible samples.

5 Static Structural Estimation

What can joint observations of testing, revision, adoption, and nonintervention tell us about editorial learning and its cost? Rare testing alone has several explanations: experiments may be costly or uninformative, challengers unattractive or already clearly superior, later audiences small, reviews infrequent, or tests imperfectly detected. We use these observations jointly under a parsimonious model rather than equating a low action frequency with a resource cost.

The static problem already contains the central investment trade-off: learning uses an experimental audience and resources to improve choices for later readers. It is a natural starting point for estimation. At each review the editor compares retaining, switching, and testing myopically. Reviews can recur after retention, but the editor does not value future reconsideration through equation (21). Estimation of that dynamic continuation problem is deferred to a subsequent draft.

5.1 Economic environment and estimable specification

In applying the model, we distinguish an article's economic audience from our ability to observe its headline. Let $H_{it} = 1$ indicate homepage display and $H_{it} = 0$ its absence. At scrape time ℓ, successful headline coverage is $C^h_{i\ell}$, and the headline is observed only when $O^h_{i\ell} = H_{i\ell}C^h_{i\ell} = 1$. Articles can continue attracting readers and undergoing editorial changes after departure, while sharing may remain observed.

The likelihood uses eligible observed histories from Section 4. Homepage departure and loss of reliable coverage censor those histories; unobserved actions supply no evidence of editorial inaction. Conditional on reference demand and current status, the baseline holds homepage-transition probabilities independent of the focal testing or switching choice. Appendix A.4 gives that dynamic specialization; the estimated static model uses the predecision expected-

audience forecasts.

The editor has an incumbent and an available, persistent challenger. We restrict the general prior in Section 3 to common payoff coefficients, a common challenger intercept, and constant prior uncertainty:

$$\Delta_i \mid d_i \sim N(\mu_i, \sigma^2), \qquad \mu_i = \alpha_\mu + \beta' d_i, \qquad s(n) = \sigma\sqrt{\frac{n}{n+\chi}}, \quad \chi = 3. \tag{25}$$

The intercept permits a prior advantage not explained by measured wording. It can reflect omitted characteristics, private judgments, or selection and need not be an unbiased quality assessment. The eight differences use the headline-only measures available in the prepared structural release. We do not separately estimate the editorial-loss component in equation (2); the available excerpt-support score is not an independently validated measure of that loss.

With discount factor one, expected audiences $(\widehat{n}, \widehat{M})$ enter the static values in equations (12)–(14). Direct switching earns $(n + M)\mu$ immediately; testing earns $n\mu/2$ during balanced allocation, then $MJ(\mu, s(n))$, less $k \geq 0$. Thus both information resolution and foregone deployment matter. The fixed ratio χ means that an experimental audience of three normalized forecast shares resolves half the prior variance. This is a maintained precision restriction, not a measured number of randomized readers.

Structural estimation includes weak first binary cycles to avoid selecting against experiments that return to the incumbent.[4] The same histories retain 4,613 direct revisions; the 4,805 broad adoption episodes do not enter as additional initiations. The 149 multi-headline first detections censor binary risk rather than supplying artificial pairs.

5.2 Editorial review and action probabilities

The fifteen-minute grid measures observation time, not newsroom meetings. Actual reviews arrive at a constant Poisson rate ρ per hour. Independent extreme-value action shocks, with

[4]Strong detection requires three transitions, potentially retaining A–B–A–B challenger victories while excluding two-transition A–B–A incumbent victories. Adding 260 weak boundaries (253 incumbent settlements and seven unresolved) gives 1,075 tests, 1,042 completed outcomes, and 758 challenger adoptions (72.74 percent). The 815 strong binary tests have 96.07 percent challenger adoption among 789 completed outcomes. Weak cycles do not verify randomization: brief challenger display and a single reversion can reflect reconsideration. Appendix E.1 reports detection and duration sensitivities.

common scale normalized to one, give

$$P_a(d, S_{it}) = \frac{\exp Q_a(d, S_{it})}{1 + \exp Q_S(d, S_{it}) + \exp Q_T(d, S_{it})}, \quad a \in \{K, S, T\}, \tag{26}$$

$$\ell_{S,it}(d) = \rho P_S(d, S_{it}), \qquad \ell_{T,it}(d) = \rho P_T(d, S_{it}). \tag{27}$$

At a review, retention produces no intervention. An unchanged interval can therefore mean no review or a review followed by retention. Its likelihood is survival under the sum of the two intervention intensities, using actual eligible duration, including shortened terminal intervals. Action shocks capture unobserved considerations affecting a particular review; the estimated review rate is not an observed meeting count.

5.3 Unobserved challengers and the structural likelihood

Unrevealed alternatives are economically consequential. An article with repeated nonintervention can have an unattractive candidate, even when its public attention state resembles that of a tested article. We integrate over that candidate once for the complete history. Redrawing after each unchanged interval would discard the information in its persistence.

The entry difference in coordinate j has a bounded truncated-normal distribution:

$$d_{ij} \mid x_{Ai0} \sim \mathrm{TN}\Big(c_j a_j, \tau^2 c_j^2;\ L_j - x_{Ai0,j},\ U_j - x_{Ai0,j}\Big). \tag{28}$$

Sentiment levels lie in $[-1, 1]$ and other levels in $[0, 1]$. The eight locations and common dispersion are estimated jointly. The fixed c_j set difference units; observed selected pairs are not treated as draws from an unselected candidate population. Coordinates are conditionally independent in this first specification. Minor incumbent changes update the contrast against the same challenger.

Let $\mathcal{H}_i(d) = \sum_t [\ell_{S,it}(d) + \ell_{T,it}(d)] e_{it}$ be cumulative intervention hazard. A censored history contributes

$$\mathcal{L}_i^C = \int g_\psi(d \mid x_{Ai0}) \exp[-\mathcal{H}_i(d)]\, dd. \tag{29}$$

A revealed challenger at action $a_i \in \{S, T\}$ contributes its density, preceding survival, and

terminal action intensity:

$$\mathcal{L}_i^a = g_\psi(d_i \mid x_{Ai0}) \exp[-\mathcal{H}_i(d_i)] \ell_{a_i,i\tau}(d_i) \left[p_i^{b_i}(1-p_i)^{1-b_i}\right]^{R_i}, \quad p_i = \Phi\left(\frac{\mu_i}{s(n_i)}\right). \tag{30}$$

Here $R_i = 1$ only for a completed test with observed adoption, and $b_i = 1$ denotes adopting the challenger. The private signal is integrated out, and unresolved tests receive no imputed winner. The initiation and adoption factors use the same pair; testing cost is sunk at adoption but affects which pairs enter experimentation.

The likelihood conditions on entry eligibility and independently censored observation given its maintained state. Strong actions retain the prepared display-change midpoint and weak boundaries their prepared lower endpoint. These operational times are not internal randomization starts. Appendix E gives the numerical implementation and observation restrictions.

5.4 Identification and estimation

Different observations constrain different parts of the same model. Selected headlines and adoption relate priors to information resolution. Survival restricts which alternatives remain private. Revisions distinguish immediate deployment from learning. At a given experimental capacity, remaining audience changes the return to information; capacity changes both precision and allocation cost. Action frequencies and durations help separate choice probabilities from review intensity under the constant-review restriction.

We normalize the common action-shock scale to one and one predicted archived share to one opportunity unit. Direct switching has zero incremental cost; the test-start cost is nonnegative and is not accompanied by a separate free test intercept. Fixing $\chi = 3$ restricts information precision because the data lack actual experimental sample sizes. Forecast uncertainty about demand cannot substitute for uncertainty about relative headline quality.

We maximize the full-history likelihood using fixed candidate draws and analytic Gaussian adoption. The reported parameters use the 1,024-draw fit underlying all subsequent counterfactuals. Conditional article-sandwich standard errors hold forecasts, scores, and event classification fixed. Numerical refinement, parameter recovery, and alternative review/prior restrictions are reported in Appendix E. Those checks address implementation and conditional recoverability; measurement and economic fit require separate evidence.

5.5 Structural estimates

Table 2 reports the central parameters. The test-start cost, 1.6213, measures the additional hurdle to acquiring information rather than switching directly. With the action-shock scale normalized to one and forecast shares as opportunity units, it is an editorial-payoff penalty.

Table 2: Preferred static parameters

Parameter	Estimate	Conditional SE
Test-start cost k	1.6213	0.0418
Prior standard deviation σ	0.0036985	0.0010216
Common prior intercept α_μ	0.0024116	0.0009840
Editorial reviews per hour ρ	0.02637	0.00107
Candidate dispersion τ	1.1591	0.0093
Signal-noise ratio χ	3 (fixed)	–
Direct-switching cost	0 (fixed)	–

Common action-shock scale is one; exposure units are expected archived URL shares. Standard errors use the matched full-history article-sandwich covariance and the delta method. Forecast, score, and detection uncertainty are excluded.

The positive prior intercept, 0.0024116, means the editor often regards an available alternative as promising even before accounting for measured attributes. Prior uncertainty, 0.0036985, leaves room for a test to reverse that judgment. A clearly favorable challenger encourages immediate revision; a close, uncertain comparison makes learning more useful. Readability has the largest positive coefficient in the expected-payoff index, while sentiment and information withholding enter negatively. The model therefore links anticipated quality to both clarity and relatively negative wording, while allowing unmeasured appeal. Appendix Table A9 reports all coefficients and the estimated distribution of available challengers.

The economic content of the cost estimate rests on these components jointly. Rare testing can reflect high cost, limited learning benefits, or infrequent review. Audience responses and adoption therefore provide checks beyond fitting the aggregate number of tests.

5.6 Model fit and the remaining-audience relationship

The fitted model closely reproduces aggregate interventions but overpredicts challenger adoption (Table 3). Its 1,051.5 tests and 4,623.3 revisions compare with 1,075 and 4,613 observed; predicted adoption is 82.74 percent against 72.74 percent on the same 1,042 completed pairs. Initiation forecasts integrate candidates using preceding survival, without

inserting the subsequently revealed challenger. Adoption predictions condition on that revealed pair.

Table 3: Observed choices and fitted static behavior

Moment	Observed	Fitted
First binary tests	1,075	1,051.5
Direct revisions	4,613	4,623.3
Challenger adoption (%)	72.74	82.74
Expanded all-test audience contrast	0.01543	0.01095
Binary testing	0.01052	0.01083
Any binary intervention	-0.00578	0.02707
Test share among interventions	0.82796	0.99808

Counts sum competing-event probabilities over actual eligible durations. Below the rule, audience contrasts are percentage-point changes from conditional Q25 to Q75. The first three concern fifteen-minute probabilities; the last concerns the test share among interventions. Binary diagnostics use identical 965,511 standardization rows. The expanded all-test comparison uses 1,061,482 rows and adds an independent multi-headline nuisance hazard to the fitted binary hazard.

Section 4.6 compares articles with different predicted later readership while keeping experimental capacity and observed states comparable. Matching the total number of tests is only part of the task. The model should also explain which articles editors choose to test, particularly when more readers remain to benefit from learning. For two equally informative thirty-minute experiments, a larger later audience increases the return to selecting the better headline. With a constant estimated review rate, this relationship comes from the relative values of testing, switching, and retention.

Matching event definitions, the expanded all-test empirical contrast is 0.01543 percentage points and the fitted contrast 0.01095, or 71.0 percent of that estimate. The fitted comparison adds a constant nuisance hazard for the 149 multi-headline detections. Restricting both comparisons to binary tests gives an observed increase of 0.01052 points and a fitted increase of 0.01083. The model thus closely reproduces this binary relationship between later audience and testing. Testing's share among interventions also rises, by 0.828 points in the data and 0.998 in the model. A shift toward learning can involve substitution away from immediate revision, so this choice between the two actions is economically informative. Table 3 and Appendix F retain the broader intervention diagnostics.

Constant reviews miss the early concentration of observed interventions. The static values already reward a larger remaining audience, but each review is myopic. The dynamic waiting decision in equation (21) additionally weighs readers lost through delay against future demand

information and testing opportunities. Estimating that continuation value may help explain when editors act. Review rates, challenger availability, and detection can also change over an article's life, so dynamic choice is one potential explanation of the timing discrepancy.

Chronological prediction provides a further assessment. Estimating the same static specification on 24,818 earlier articles and predicting 11,416 later articles yields 369.9 tests against 162 observed and 82.77 percent adoption against 56.41 percent. Later revisions are underpredicted. Appendix F reports these results and selected-candidate comparisons. The counterfactuals use the full-sample estimates, with their substantive adoption and timing errors.

6 Counterfactual Analysis

How would the headlines readers encounter differ without experimentation or direct revision? We remove actions from the fitted editor's menu and recompute choices, keeping the initial article states, available challengers, priors, and reference audience fixed. These exercises isolate deployment and selection; they do not allow journalists to change which challengers they produce.

6.1 Counterfactual editorial regimes

We compare keeping the initial headline throughout; allowing direct revision alone; allowing testing alone; and allowing both. Testing alone prohibits immediate switching but permits adoption of the challenger after a favorable experimental result. Removing an action renormalizes the logit probabilities over the remaining menu. It therefore changes both the frequency and composition of intervention, rather than merely setting one action's probability to zero.

The main trajectories cover 27,766 articles with a modern forward-forecast vintage and entry age at most twenty-four hours. These restrictions keep rolling forecasts within the fitted seventy-two-hour age support. The other 8,468 estimation articles retain valid single-review comparisons but lack this common trajectory support. Table 4 reports outcomes over a common twenty-four-hour evaluation window after entry. The editor's decision values still use a rolling twenty-four-hour audience at each review. A late action may consequently have benefits beyond the outcome window.

To compare these options, we follow the same articles under each regime. Each starts with

its observed headline, a possible alternative drawn from the estimated distribution, and a forecast of future attention. We hold that alternative and the underlying audience conditions fixed across regimes, changing only the available actions. Over the next twenty-four hours, the editor can reconsider the headline at subsequent review opportunities until an intervention begins. Direct revision implements the alternative immediately; testing compares the pair and then retains the headline favored by its result. After revision or test completion, the settled headline remains in place for the evaluation.

Forecasts advance using expected sharing rather than actual future tweets, with a common reference audience path across regimes. Reviews can continue after homepage departure. At each review the editor follows the estimated static choice rule, without valuing the option of a future reconsideration. These repeated choices therefore describe the current static model rather than optimal dynamic waiting. Appendix G details the forecast construction and numerical calculation.

Table 4: Editorial options over a common twenty-four-hour horizon

Outcome	Keep	Revision	Testing	Both
Direct revision (%)	0.000	31.120	0.000	27.100
Testing (%)	0.000	0.000	14.495	5.865
Permanent challenger adoption (%)	0.000	31.120	11.923	31.945
Final negativity	0.256271	0.275929	0.263941	0.276752
Exposure negativity	0.318727	0.325026	0.321996	0.324921
Final classifier extremity	0.363349	0.381895	0.370093	0.382571
Exposure classifier extremity	0.397382	0.402614	0.399620	0.402713
Gross quality-selection payoff	0.000000	0.108181	0.083573	0.116921
Test-start cost	0.000000	0.000000	0.235007	0.095085
Net systematic payoff	0.000000	0.108181	-0.151435	0.021836

27,766 articles; all parameters and candidate generation held fixed. Revision and Testing permit retention and the named initiation action; adoption of a test winner remains possible whenever testing is available. Final scores weight articles equally; exposure scores weight forecast opportunities. Payoffs are normalized per article relative to retaining the initial headline.

With both options, 5.865 percent of articles receive a test and 27.100 percent a direct revision. Removing testing raises revision to 31.120 percent; removing revision raises testing to 14.495 percent. These are simulated full-horizon frequencies, distinct from detected events over the shorter, censored estimation histories. Adding testing when revision is possible therefore differs from adding it to a keep-only menu: it also displaces immediate deployment.

6.2 Effects on headline characteristics

We also report two transformations of the headline attribute scores. Negativity is the negative part of signed sentiment, $\max\{0, -x_{\text{sentiment}}\}$. Classifier extremity is $2|p_D - 0.5|$, where p_D is Democratic-associated language probability. The latter measures distance from the classifier's neutral point, not validated ideological extremism or political polarization. We transform each candidate and incumbent before averaging; a nonlinear score of an average wording vector would be a different object.

Table 5 compares adding testing to a regime that already allows revision with adding revision when testing is available. Relative changes below use the score in the regime receiving the additional option; Appendix Table A16 gives those reference levels.

Table 5: Headline-presentation changes across news sections

	Negativity	Democratic language	Readability
A. Adding testing to direct revision			
	Final / exposure-weighted: $\Delta \times 1{,}000$; [% change]		
All articles	+0.823 / -0.105	+0.157 / +0.236	+0.257 / +0.113
	[+0.298% / -0.032%]	[+0.032% / +0.044%]	[+0.034% / +0.015%]
U.S. news	+0.544 / -0.409	+0.201 / +0.553	-0.005 / +0.412
	[+0.175% / -0.134%]	[+0.040% / +0.096%]	[-0.0007% / +0.050%]
World	+0.431 / -0.219	+0.115 / +0.100	+0.012 / -0.161
	[+0.120% / -0.057%]	[+0.023% / +0.020%]	[+0.0015% / -0.021%]
Opinion	+0.498 / +0.346	+0.164 / +0.110	+0.559 / -0.132
	[+0.141% / +0.082%]	[+0.033% / +0.021%]	[+0.088% / -0.022%]
B. Adding direct revision to testing			
	Final / exposure-weighted: $\Delta \times 1{,}000$; [% change]		
All articles	+12.811 / +2.925	+1.146 / -0.128	+1.895 / +1.187
	[+4.854% / +0.908%]	[+0.231% / -0.024%]	[+0.250% / +0.156%]
U.S. news	+12.498 / +3.613	+0.575 / -0.670	-4.787 / -0.699
	[+4.193% / +1.200%]	[+0.113% / -0.116%]	[-0.578% / -0.085%]
World	+6.673 / +1.287	+1.612 / +0.353	-1.877 / +0.296
	[+1.883% / +0.338%]	[+0.328% / +0.070%]	[-0.235% / +0.038%]
Opinion	+4.869 / -0.188	+1.568 / -0.193	+14.924 / +6.981
	[+1.393% / -0.045%]	[+0.320% / -0.037%]	[+2.395% / +1.171%]

Each first row gives final / exposure-weighted score changes multiplied by 1,000. The bracketed row below gives percentage changes from the corresponding revision-only (Panel A) or testing-only (Panel B) score, using unrounded values. Political-language percentages are proportional changes in the classifier probability, not polarization or ideological extremity; raw score differences remain the primary direction measure. Final means weight articles equally; exposure means use forecast audience weights within each population. All articles: 27,766; U.S.: 5,808; World: 4,296; Opinion: 4,049. Parameters and candidates are common across menus.

Adding testing to revision raises final negativity by about 0.30 percent while lowering exposure negativity by 0.03 percent. Adding revision to testing produces larger increases of 4.85 and 0.91 percent, respectively. These comparisons use the score in the regime receiving the new option, rather than retention as a universal baseline. Experimentation's marginal effect on average headline characteristics is small, partly because relatively few articles are tested and some tests replace immediate revisions.

Final selection and reader exposure also differ within sections. Testing raises final negativity in U.S. news, World, and Opinion, but lowers exposure negativity in the first two and raises it in Opinion. Each change is below 0.2 percent of its revision-only score. Revision produces larger final changes, yet Opinion's exposure negativity falls slightly. The selected headline and the amount of audience reached by it are separate parts of the comparison.

Readability and negativity need not move together. Adding revision raises final negativity in all three sections, while increasing readability in Opinion and lowering it in U.S. and World news. Opinion's final readability increase is about 2.4 percent, substantially larger than testing's small changes. Editors choose bundles of attributes, so other qualities can offset a clarity advantage. Appendix G.1 retains the other headline characteristics.

Partisan-direction changes are small in score points. Testing moves final and exposure direction toward Democratic-associated language in all three sections. Revision raises final direction but moves exposure toward Republican-associated language in U.S. news and Opinion and toward Democratic-associated language in World. The proportional classifier-score changes in brackets provide scale; they do not measure changes in distance from political neutrality.

All sections share the fitted payoff coefficients. Their stories, initial headlines, feasible alternatives, and audience paths differ, changing both which articles intervene and how long the selected headline is shown. Section 4.3.1's regressions describe attribute selection among observed tested pairs; the counterfactuals instead ask how access to an editorial option changes displayed headlines. They neither estimate the same quantity nor imply separate structural preferences by section.

6.3 Exposure-weighted headline presentation

A final headline can overstate an editorial option's reach. If 1,000 readers arrive before a late change and 100 afterward, most see the earlier headline even when the final headline

differs substantially. We therefore weight the incumbent before intervention, temporary half-challenger exposure during testing, and the selected headline after revision or settlement by a common forecast audience. The denominator is identical across regimes.

The distinction changes the testing comparison (Figure 7). Adding testing to revision changes exposure negativity by -0.000105, despite increasing final negativity; its conditional parameter band is $[-0.000337, 0.000077]$. Exposure classifier extremity rises by only 0.000099, with band $[-0.000042, 0.000184]$. Testing delays settlement and substitutes for some immediate revisions, so a more negative final selection need not mean more negative average exposure.

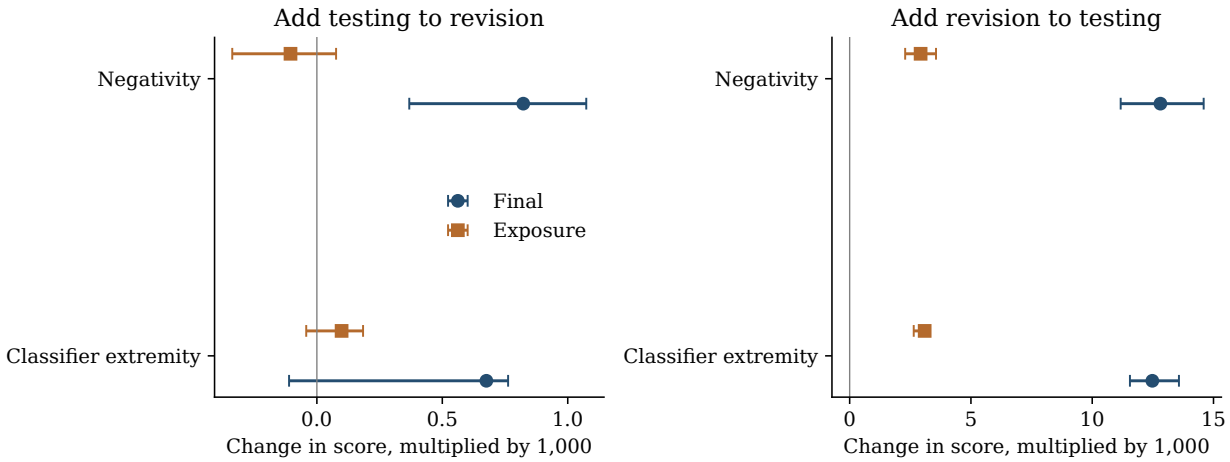


Figure 7: Final and exposure-weighted headline characteristics respond differently

Score changes multiplied by 1,000. Points compare adding one option when the other is available. Bars are conditional 95 percent structural-parameter bands using 128 matched article-sandwich draws; forecasts, scoring, detection, and candidate-family restrictions are held fixed. Classifier extremity is distance from political-language neutrality, not an ideological-polarization measure.

Adding revision to testing has larger audience-weighted effects: negativity rises by 0.002925 and classifier extremity by 0.003092. Direct revision can implement a favorable prior judgment immediately, whereas testing uses time before a final choice. Exposure readability and specificity also rise. Under both options, 85.57 percent of reference exposure remains with the incumbent, 0.112 percent is temporary challenger exposure, and 14.32 percent is permanent challenger exposure. These are modeled exposure opportunities, not observed reader assignments. Appendix G gives exposure accounting and sensitivity to review support and adoption.

6.4 Implications for editorial payoff

The preceding comparisons ask how editorial options change observable presentation. The theory asks a further question: does learning improve quality selection according to the editor's objective, and is that benefit worth its cost? Two similarly scored headlines can differ in unmeasured appeal. Testing can identify the better one without producing a large change in any recorded attribute. Post-test quality therefore enters as $\mathbb{E}[\Delta \mathbf{1}\{m > 0\}] = J(\mu, s)$, incorporating selection on the experimental result.

Adding testing to revision raises gross normalized quality-selection payoff by 0.008741 per article. Expected test-start costs rise by 0.095085, giving a net systematic-payoff change of -0.086344 (Table 4). Gross payoff already incorporates experimental allocation and deployment timing; the start cost is subtracted once.

Under this accounting, better quality selection does not cover the additional test-start cost over the evaluation horizon. The net figure excludes action-specific random utilities and benefits outside that horizon, so it is not total welfare. These counterfactuals distinguish the value of information from changes in measured headline characteristics: learning can improve headline selection with little movement in the eight scores, but its net return depends on acquisition costs and the audience available during and after experimentation.

7 Conclusion

Finite readership makes headline learning an economic decision: readers used to compare alternatives cannot benefit from the completed test. Direct revision implements an existing judgment immediately. Our Bayesian framework separates these choices and allows waiting for better information or experimental opportunities.

NYT histories reveal testing, revision, nonintervention, and adoption. Joint static estimation relates them to prior beliefs, uncertainty, review frequency, and the cost of starting a test. Beyond matching aggregate frequencies, the model generates much of the increase in testing when more readers remain afterward. This audience response connects the estimates to learning incentives; adoption and timing remain imperfectly matched.

Access to testing and direct revision changes editorial selection through different mechanisms. Revision deploys an anticipated improvement immediately, while experimentation acquires information before choosing for later readers. Final headline selection and exposure-

weighted presentation can therefore move differently. Effects on negativity, partisan-language direction, and readability vary across U.S. news, World, and Opinion under common payoff parameters. Better selection under the editor's objective need not produce large movements in measured attributes, and its net value depends on the costs of learning as well as the audience that can benefit.

Dynamic structural estimation is the next stage. Measuring the continuation value of waiting may help explain early intervention timing by weighing readers lost through delay against better future information and testing opportunities.

References

Athey, Susan and Michael Luca. 2019. "Economists (and Economics) in Tech Companies." *Journal of Economic Perspectives* 33 (1):209–230.

Aubin Le Quéré, Marianne and J. Nathan Matias. 2025. "When Curiosity Gaps Backfire: Effects of Headline Concreteness on Information Selection Decisions." *Scientific Reports* 15 (1):994.

Azevedo, Eduardo M., David Mao, José Luis Montiel Olea, and Amilcar Velez. 2023. "The A/B Testing Problem with Gaussian Priors." *Journal of Economic Theory* 210:105646.

Cochrane, Emily, Jim Tankersley, and Jeanna Smialek. 2020. "Emergency Economic Rescue Plan in Limbo as Democrats Block Action." *The New York Times* (March 22).

Ecker, Ullrich K. H., Stephan Lewandowsky, Ee Pin Chang, and Rekha Pillai. 2014. "The Effects of Subtle Misinformation in News Headlines." *Journal of Experimental Psychology: Applied* 20 (4):323–335.

Feit, Elea McDonnell and Ron Berman. 2019. "Test & Roll: Profit-Maximizing A/B Tests." *Marketing Science* 38 (6):1038–1058.

Gentzkow, Matthew and Jesse M. Shapiro. 2006. "Media Bias and Reputation." *Journal of Political Economy* 114 (2):280–316.

———. 2010. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica* 78 (1):35–71.

———. 2011. "Ideological Segregation Online and Offline." *Quarterly Journal of Economics* 126 (4):1799–1839.

Hagar, Nick and Nicholas Diakopoulos. 2019. "Optimizing Content with A/B Headline Testing: Changing Newsroom Practices." *Media and Communication* 7 (1):117–127.

Iyengar, Shanto. 1991. *Is Anyone Responsible? How Television Frames Political Issues*. Chicago: University of Chicago Press.

Iyengar, Shanto and Kyu S. Hahn. 2009. "Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use." *Journal of Communication* 59 (1):19–39.

Jensen, Jakob D. 2008. "Scientific Uncertainty in News Coverage of Cancer Research: Effects of Hedging on Scientists' and Journalists' Credibility." *Human Communication Research* 34 (3):347–369.

Keller, Godfrey and Sven Rady. 1999. "Optimal Experimentation in a Changing Environment." *Review of Economic Studies* 66 (3):475–507.

Koning, Rembrand, Sharique Hasan, and Aaron Chatterji. 2022. "Experimentation and Start-up Performance: Evidence from A/B Testing." *Management Science* 68 (9):6434–6453.

Leung, Tin Cheuk and Koleman S. Strumpf. forthcoming. "Measuring Media Bias Through Homepage Duration: Evidence from NYT and WSJ." *Journal of Law & Economics* .

Loewenstein, George. 1994. "The Psychology of Curiosity: A Review and Reinterpretation." *Psychological Bulletin* 116 (1):75–98.

Petty, Richard E., John T. Cacioppo, and Rachel Goldman. 1981. "Personal Involvement as a Determinant of Argument-Based Persuasion." *Journal of Personality and Social Psychology* 41 (5):847–855.

Prior, Markus. 2013. "Media and Political Polarization." *Annual Review of Political Science* 16:101–127.

Robertson, Claire E., Nicolas Pröllochs, Kaoru Schwarzenegger, Philip Pärnamets, Jay J. Van Bavel, and Stefan Feuerriegel. 2023. "Negativity Drives Online News Consumption." *Nature Human Behaviour* 7:812–822.

Rothschild, Michael. 1974. "A Two-Armed Bandit Theory of Market Pricing." *Journal of Economic Theory* 9 (2):185–202.

Rust, John. 1987. "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher." *Econometrica* 55 (5):999–1033.

Shen, Yikang. 2026. "Platforms as Editors: How Algorithmic Curation Shapes Online News Slant." Working paper, Carnegie Mellon University. September 30 draft.

Shulman, Hillary C., David M. Markowitz, and Todd Rogers. 2024. "Reading Dies in Complexity: Online News Consumers Prefer Simple Writing." *Science Advances* 10 (23):eadn2555.

A Theory derivations

A.1 Gaussian updating and information value

Prior and signal precisions add, so $v^+ = (v^{-1} + n/\nu^2)^{-1}$. Since $\mathrm{Var}(Z) = v + \nu^2/n$,

$$\mathrm{Var}\{m(Z)\} = \left(\frac{v}{v+\nu^2/n}\right)^2 (v+\nu^2/n) = \frac{v^2}{v+\nu^2/n} = v - v^+. \tag{A1}$$

For $U \sim N(0,1)$ and $s > 0$,

$$J(\mu, s) = \int_{-\mu/s}^{\infty} (\mu + su)\phi(u)\, du \tag{A2}$$

$$= \mu\Phi(\mu/s) + s\phi(\mu/s), \tag{A3}$$

using $\int_a^\infty u\phi(u)du = \phi(a)$. Subtracting μ^+ and applying normal symmetry gives equation (10). Differentiation cancels density-derivative terms and yields equation (11). Convexity of the maximum also establishes $G \geq 0$. At $\mu = 0$, $G = s/\sqrt{2\pi}$; as $s \to 0$, $J \to \mu^+$.

The ex ante adoption probability is $\Phi(\mu/s)$. More information moves this probability toward one half, from above or below depending on the prior mean. Its effect on the chance of choosing B need not be positive even though decision value rises. Quality of information is therefore not measured by challenger-adoption frequency alone.

A.2 Static surplus and sample size

With $N = n + M$, subtracting best immediate deployment gives

$$Q_T - N\mu^+ = \frac{n\mu}{2} + (N-n)(G+\mu^+) - k - N\mu^+ \tag{A4}$$

$$= (N-n)G - \frac{n|\mu|}{2} - k. \tag{A5}$$

This proves Proposition 1. At fixed n, each additional post-test exposure contributes G; $\partial M^*/\partial k = 1/G$ when $G > 0$. If $M = 0$, information has no deployment value.

Writing $\chi = \nu^2/v$,

$$s(n) = \sqrt{v}\sqrt{\frac{n}{n+\chi}}, \qquad s'(n) = \frac{\sqrt{v}\,\chi}{2\sqrt{n}(n+\chi)^{3/2}}. \tag{A6}$$

At fixed total audience, an interior positive-size test satisfies

$$(N-n)\phi(\mu/s(n))s'(n) = G(\mu, s(n)) + |\mu|/2. \tag{A7}$$

The gain from more information equals lost deployment value plus experimental allocation cost. A fixed k changes whether to test, not this conditional first-order condition. At prior indifference,

$$2n^2 + 3\chi n - \chi N = 0, \qquad n^* = \frac{\sqrt{9\chi^2 + 8\chi N} - 3\chi}{4}. \tag{A8}$$

This candidate optimum must still beat no testing after subtracting k. The fixed-clock baseline instead maps traffic into n.

A.3 Timing and economic continuation

Suppose audience deterministically shrinks, the same fixed-sample experiment remains feasible, beliefs/precision do not change, no information arrives while waiting, and the editor commits after intervention. Without discounting,

$$Q_T(t;k) - Q_T(t+\ell;k) = (N_t - N_{t+\ell})J(\mu, s(n)) \geq 0. \tag{A9}$$

Test-stage payoff and k cancel. For $\mu > 0$, immediate switching likewise dominates delayed switching by $(N_t - N_{t+\ell})\mu$; for $\mu \leq 0$, retention is preferable. Thus each eventual intervention is weakly better now. A constant positive cost alone cannot produce strict delay in this benchmark.

Fixed-clock traffic variation breaks the common-sample comparison; demand learning with $k > 0$ can additionally avoid low-value tests. Recurrent options or action-dependent placement change continuation and require their own comparison.

A.4 Homepage specialization of the reference audience process

For the NYT application, write the main-text audience state as $\mathcal{X}_t = (X_t, H_t)$. Here $H_t = 1$ indicates homepage display and $H_t = 0$ its absence; X_t contains other reference-demand conditions, including age, calendar time, status-spell age, topic conditions, and beliefs about

attention. The editor knows the transition probabilities, but not their future realizations:

$$p_{hh'}(X_t) = \Pr(H_{t+1} = h' \mid H_t = h, X_t), \qquad h, h' \in \{0, 1\}. \tag{A10}$$

Conditional on current status and reference demand, testing or switching leaves these probabilities unchanged. Transition shocks are independent of private candidate quality and the experimental signal. Relative headline payoffs are common across states. Departure can reduce traffic without ending readership or experimentation, and return remains possible. Appendix H relaxes action-independent placement.

The components of equation (1) partition exposure by homepage status when it occurs, rather than referral source:

$$F_{it}^h = \mathbb{E}_t \left[\sum_{j=0}^{T_i - t - 1} \delta^j \mathbf{1}\{H_{i,t+j} = h\} r_{i,t+j} \right], \quad h \in \{0, 1\}, \tag{A11}$$

$$F_{it} = F_{it}^1 + F_{it}^0. \tag{A12}$$

Both components can be positive. Similarly, n^h and M^h count experimental and later exposure in state h, giving $n = n^1 + n^0$ and $M = M^1 + M^0$. For a Markov-sufficient reference state,

$$F_t(X, h) = \bar{r}_t(X, h) + \delta \sum_{h'=0}^{1} p_{hh'}(X) \mathbb{E}[F_{t+1}(X', h') \mid X, h, h']. \tag{A13}$$

Here $\bar{r}_t(X, h)$ is expected current exposure and X' is next period's non-placement reference state. By the law of total expectation, the waiting value in equation (20) becomes

$$Q_W(S_t; k) = \delta \sum_{h'=0}^{1} p_{H_t h'}(X_t) \mathbb{E}[V_{t+1}((X', h'), d_i, \mu_t, v_t; k) \mid X_t, H_t, h'] \,. \tag{A14}$$

Both state continuations appear and $\bar{r}_t(X, 0)$ may be positive. Test completion uses $F_{t+L_T} J$ without requiring homepage presence. Terminal value is set at economic termination, not departure or the last scrape. Observation and censoring restrictions enter the estimated histories as described in Section 5; they do not terminate the economic problem.

B Data construction and headline measurement

B.1 Event detection and risk populations

Classification uses repeated alternation, distinct headlines, and thirty minutes of subsequent settlement within spells broken by gaps exceeding five minutes. Strong first tests require at least three transitions. Completed adoption instead retains A–B–A reversions, avoiding selection against incumbent victories. The 4,805 adoption episodes therefore differ from the 964 strong first tests in the hazard sample. All 484 incomplete two-/multi-way episodes have missing winners; a later headline is not imputed as immediate adoption.

The classifier's first-seventy-two-hour histories cover 53,316 of 56,714 archive articles. Conditional on that narrower universe, revision and testing frequencies are 16.56 and 7.54 percent, versus 15.57 and 7.09 using the main archive denominator. The earliest classified intervention can follow an unresolved sequence and is not necessarily a lifetime first action. Among first classified two-way tests and revisions, median publication ages are 4.91 and 4.83 hours; among hazard-eligible revisions and strong tests they are both about 4.64 hours. None of these event distributions is a rate per eligible risk hour. The timing audit and complete event exports remain in the reproducibility archive.

Table A1 traces risk construction. Events truncate exposure at the onset bracket's midpoint; weak first tests censor the strong-test history rather than supplying a full no-event interval. The ten-minute incumbent-stability rule excludes 2,070 articles in an otherwise-identical marginal comparison, or 4.24 percent of articles eligible without it. Sequential attrition counts condition on different preceding restrictions and should not be read as that marginal comparison. The final panel contains 340,846.76 eligible risk hours. Date-intercept likelihoods retain 943,800 informative testing intervals and 1,363,877 revision intervals after removing dates with no event of the relevant type.

B.2 Descriptive attention profiles

The processed article-hour panel supplies all seventy-two publication-relative hours for each of the 56,714 homepage-linked articles with a valid publication timestamp. We retain all hours, including 2,367,347 with no archived shares, without requiring stable wording or continued homepage display. Their 4,083,408 article-hours contain 21,605,755 recorded shares.

Table A1: Construction of the decision-risk population

Restriction	Articles	Rows
Homepage records	56,714	61,546,862
Candidate 15-minute grid	53,316	3,157,085
First spell	53,316	2,555,507
Known state and coverage	52,193	2,495,834
No active experiment	52,193	2,480,776
Usable incumbent wording	51,444	2,443,405
Ten-minute entry age	48,145	2,392,109
Eligible risk panel	46,790	1,775,638
Chronological forecast available	37,676	1,452,627
Publication age $\leq$ 48h	36,234	1,380,937
Positive eligible exposure	36,234	1,380,928

Rows are archive records at the first step and review-grid observations thereafter. The final population includes 964 strong tests and 4,613 revisions. No future completion restriction determines risk entry.

Publication-age means divide each hourly count by 56,714. This descriptive universe differs from both the broader tweet archive and the stable-wording response sample.

For clock-time denominators, each publication-relative hour is split at UTC clock-hour boundaries and fractional durations are accumulated at their New York local hour. The already processed share timestamps supply exact local-hour numerators, reconciling to the hourly-panel total; raw API records are not reprocessed. Timezone-aware conversion to `America/New_York` preserves repeated fall hours and omitted spring hours. Within every clock hour we calculate rates separately for all seventy-two publication-age hours. Their equally weighted average gives the age-adjusted line. Thus clock profiles have a common age composition, rather than reflecting the number or youth of articles published at particular times. Adjustment is descriptive and does not hold all story characteristics fixed. A zero remains absence of an archived record, not verified complete acquisition.

B.3 Headline scores

The frozen dictionary contains 80,271 article–raw-headline combinations, representing 77,823 distinct natural texts and 55,644 articles. Text states are distinguished before scoring. Sentiment uses the positive-minus-negative probability from the pinned Twitter-RoBERTa model, retaining negation and natural wording.

Partisan training starts from 4,204,058 raw records, including 4,142,719 linked to 2,100 Democratic/Republican accounts. Politician identities are linked before splitting so campaign

and official accounts cannot leak across evaluation groups. Removing institutional accounts, duplicates, and conflicting text yields 275,770 tweets from 1,509 groups. The calibrated logistic model uses TF–IDF word unigrams/bigrams and character three- to five-grams. Held-out-politician accuracy is 72.34 percent and ROC area 0.8078. Transfer calibration uses shared titles with purged overlap; evaluation on 22,377 titles shared by unseen politicians gives accuracy 65.07 percent and ROC area 0.7117. These tests predict party labels, not independently coded NYT bias.

The adoption score retains the established context-assisted specification, with a gate assigning 0.5 where supported U.S.-political evidence is absent. Its metadata snapshot is not certified as historically available. Hazard controls therefore use the headline-only counterpart. Limited variation around neutrality is consequential for some section comparisons, especially Sports.

The six additional scores use pinned `Qwen3-30B-A3B-Instruct-2507` inference with a separate fixed rubric for each attribute (Table A2). Inputs exclude adoption, clicks, section, and the other scores. If $\pi_{ha\ell}$ is the normalized next-token probability of category $a \in \{0, \ldots, 4\}$,

$$x_{h\ell} = \frac{1}{4} \sum_{a=0}^{4} a\pi_{ha\ell}, \qquad \sum_{a=0}^{4} \pi_{ha\ell} = 1. \tag{A15}$$

Category probabilities use temperature one; normalization uses float64 with endpoint-roundoff clipping. The transformation uses no empirical quantiles. A score measures the model's assessment against fixed anchors, not calibrated human judgments or overall quality.

Only withholding may use an article excerpt. The latest body recorded at or before first headline observation is frozen for recurrences. Bodies require eighty words; selected paragraphs are capped at 360 words. Context is used only when a separate DeBERTa textual-entailment screen reaches 0.5; otherwise the headline-only measure remains. There are 43,104 variants with recorded bodies and 27,954 with preferred contextual withholding. Textual support from an excerpt is not verified factual accuracy. The five other dimensions use wording alone.

Prompt, punctuation, ordinal-temperature, and inference-cache checks support implementation consistency; the preferred rubric passes twelve synthetic directional comparisons. They are not independent human validation. Recorded diagnostics include false withholding/support classifications, endpoint concentration, inherited title-identity contamination,

Table A2: Fixed ordinal anchors for the six additional attributes

Dimension	0	1	2	3	4
Information withholding	Direct central development; no material teaser	Mostly direct; small gap	Main development given; meaningful central gap	Substantial promised answer withheld	Essential answer chiefly withheld to invite reading
Readability	Meaning very difficult to recover	Considerable jargon or ambiguity	Understandable with effort	Largely clear; minor complexity	Immediately understandable and direct
Specificity	No identifiable development	Broad topic or vague development	Some concrete facts; central elements broad	Clear, informative development	Highly concrete central facts
Personal consequences	No tangible personal consequence	Weak or indirect implications	Affected group or consequence indicated	Clear substantial effect on people's lives	Concrete lived consequences foregrounded
Agency / responsibility	No responsible actor	Responsibility weakly implied	Actor or responsibility partly identified	Actor identified; responsibility indirect or shared	Responsible actor and action explicitly attributed
Linguistic certainty	Outcome open or no substantive assertion	Tentative possibility	Qualified, conditional, or alleged proposition	Mostly asserted; limited qualification	Categorical claim or definite prediction

Condensed definitions from the implemented rubrics. Higher scores indicate more of the named attribute; they do not rank overall headline quality. Withholding excludes ordinary compression and genuinely unresolved news. Readability does not reward shortness alone, specificity does not count names mechanically, personal consequences require explicit human implications, and agency distinguishes responsible from affected actors. Certainty measures assertiveness, not truth. Categories map to 0, 0.25, 0.5, 0.75, and 1 before probability weighting.

and possible retrospective training familiarity. A blinded human assessment remains necessary before interpreting small score differences as precise constructs. Full prompts, classifier audits, score distributions, and manual-review files remain in the research archive.

C Headline adoption and section heterogeneity

Table A3 consolidates all 64 full-sample section effects with untruncated confidence intervals. Arts/entertainment combines Arts, Movies, Theater, and Books; Other combines Technology and remaining groups. Models interact all attributes with section and omit the redundant

opinion indicator. Main-text Figure 5 compares full and first-intervention estimates. For partisan direction, first-intervention estimates per 0.1 score unit are -1.3936 points in U.S. news $[-3.9754, 1.1882]$, 4.7492 in World $[-3.7630, 13.2615]$, and 1.0761 in Opinion $[-1.8558, 4.0080]$. All intervals include zero. These are nominal article-clustered intervals; the full-sample U.S. partisan effect has BH-adjusted $q = 0.2066$ across 64 effects.

Joint section heterogeneity has $p = 0.0004$ in all episodes, $p = 0.146$ among first interventions, and $p = 0.0007$ under strong cycling. The latter sample's 87.85 percent challenger-adoption rate reflects selection against rapid incumbent reversions. Table A4 preserves the differing attribute-level evidence rather than treating the restrictions as interchangeable.

Political tags identify 1,722 political and 3,069 nonpolitical episodes; fourteen lack that classification. The overall partisan effect in political stories is -0.12 $[-2.10, 1.86]$. Within 591 U.S.-political episodes it is -3.53 $[-6.48, -0.58]$, but adjusted $q = 0.102$ and the political interaction has $p = 0.204$. Keyword classification gives joint attribute-interaction $p = 0.055$, yet partisan-interaction $p = 0.933$ and no individual interaction survives false-discovery adjustment. These results do not establish systematic partisan intensification.

Additional selection checks leave substantive uncertainty. In 767 episodes with common pre-onset body evidence, the textual-support effect is 0.59 points $[-1.34, 2.52]$, not a verified correction effect. Among 1,376 episodes with strictly pre-onset forecasts, joint attribute-by-capacity and attribute-by-remaining-attention tests have $p = 0.1690$ and $p = 0.3943$. Full sample, support, and first-intervention outputs are preserved outside the compiled appendix.

Table A3: All eight attribute associations by section

Section (episodes)	Sentiment	Partisan direction	Withholding	Readability
U.S. (1,045)	1.16 [-0.21, 2.52]	-2.75 [-5.47, -0.04]	0.81 [-0.75, 2.37]	2.18 [-0.45, 4.82]
World (644)	-1.46 [-3.17, 0.26]	10.11 [-0.07, 20.29]	1.19 [-0.96, 3.35]	2.78 [-0.64, 6.20]
Opinion (1,649)	-0.51 [-1.09, 0.08]	1.91 [-0.58, 4.40]	0.55 [-1.07, 2.18]	2.04 [0.29, 3.78]
Business (335)	-0.57 [-2.62, 1.47]	-2.51 [-12.93, 7.90]	1.14 [-1.49, 3.76]	4.30 [0.14, 8.45]
New York (246)	-2.47 [-4.71, -0.22]	0.80 [-8.42, 10.03]	-2.82 [-6.04, 0.40]	-2.58 [-7.62, 2.45]
Arts/entertainment (177)	-0.40 [-3.03, 2.23]	-4.61 [-30.14, 20.92]	0.78 [-4.27, 5.84]	-1.97 [-8.64, 4.69]
Sports (195)	-1.02 [-3.52, 1.48]	26.31 [-89.02, 141.65]	-3.56 [-6.82, -0.30]	-7.38 [-12.43, -2.34]
Other (514)	-0.16 [-1.84, 1.52]	5.26 [-3.51, 14.02]	-1.69 [-4.12, 0.74]	3.94 [0.51, 7.36]

Section (episodes)	Specificity	Personal consequences	Agency	Certainty
U.S. (1,045)	4.94 [1.87, 8.01]	-1.00 [-3.81, 1.81]	0.29 [-1.05, 1.64]	-1.58 [-3.25, 0.09]
World (644)	-0.14 [-3.76, 3.47]	-0.73 [-3.87, 2.42]	-0.39 [-1.91, 1.13]	1.70 [-0.29, 3.68]
Opinion (1,649)	-2.80 [-4.79, -0.82]	0.44 [-1.14, 2.02]	1.19 [0.11, 2.27]	-0.06 [-0.72, 0.61]
Business (335)	4.79 [0.38, 9.21]	-4.14 [-7.87, -0.42]	0.98 [-1.03, 2.99]	0.51 [-2.45, 3.48]
New York (246)	-0.66 [-6.72, 5.41]	3.19 [-1.98, 8.35]	0.51 [-2.45, 3.46]	-0.67 [-3.58, 2.24]
Arts/entertainment (177)	4.95 [-0.00, 9.90]	0.45 [-5.63, 6.53]	3.62 [-0.29, 7.52]	-4.08 [-7.66, -0.50]
Sports (195)	6.15 [0.31, 11.99]	3.23 [-1.13, 7.60]	-1.46 [-3.71, 0.78]	-1.48 [-4.65, 1.68]
Other (514)	-2.24 [-6.33, 1.86]	0.83 [-2.38, 4.04]	0.50 [-1.55, 2.56]	0.78 [-1.08, 2.64]

Average marginal effects in adoption percentage points per 0.1-unit difference; article-clustered 95 percent intervals below each estimate. All 4,805 completed episodes enter the joint section-interacted model. The exceptionally wide Sports partisan interval reflects limited score support and is displayed in full. No individual effect survives adjustment across all 64 effects.

Table A4: Attribute-specific section heterogeneity

Attribute	All: p / q	First: p / q	Cycling: p / q
Sentiment	0.1914 / 0.2552	0.2184 / 0.4369	0.1064 / 0.1774
Partisan direction	0.1123 / 0.1796	0.5791 / 0.6619	0.7630 / 0.7630
Withholding	0.0995 / 0.1796	0.4699 / 0.6266	0.1109 / 0.1774
Readability	0.0132 / 0.0527	0.0047 / 0.0377	0.0011 / 0.0090
Specificity	0.0001 / 0.0008	0.0907 / 0.2420	0.0422 / 0.1125
Personal consequences	0.2399 / 0.2742	0.0094 / 0.0377	0.2889 / 0.3852
Agency/responsibility	0.3473 / 0.3473	0.3941 / 0.6266	0.6088 / 0.6958
Certainty	0.0901 / 0.1796	0.7578 / 0.7578	0.0255 / 0.1021

Each test compares an attribute across eight sections. Benjamini–Hochberg q adjusts across eight attribute tests within each sample, separately from 64-effect and 224-pairwise adjustments. First interventions contain 2,692 distinct articles; strong cycling has 3,258 episodes from 2,730 articles.

D Attention forecasts and the remaining-audience comparison

The production framework uses nine publication-age regimes and nonoverlapping windows 0–.5, .5–1, 1–3, 3–6, 6–12, and 12–24 hours. Quarter-hour increments are nonnegative and additive, with identical baseline offsets at fitting and deployment. A/C designs use own/baseline information or add related-topic conditions. F0 updates cumulative popularity and recent own sharing; F1 adds rates from strictly earlier 0–15, 15–60, and 60–180 minute windows, including latest-minus-preceding momentum. F2 also uses observation state; these are predictive features, not causal placement effects. Recorded tags are assumed available at publication, although stored creation dates do not independently verify this.

Five expanding vintages yield 35,932 common held-out origins from 6,639 articles. Model selection uses earlier matured one-, twelve-, and twenty-four-hour losses within age regimes; the first vintage defaults to F0. The preferred policy selects F0/F1, while StateSelected permits F2. Neither editorial outcomes nor the full-sample ranking selects the preferred policy. The publication/first-fifteen-minute arrival correction remains in production: its first-five-minute mean is 0.91 shares against 0.80 observed, compared with 2.42 before correction.

The first-spell display target stops at the apparent first observation boundary and is capped by total sharing; its complement is not verified off-homepage traffic. These historical-policy predictions remain distinct from untreated exposure.

Table A5: Chronological forecast validation

Forecast	1h loss	12h loss	24h loss	24h mean
F0	9.69	126.93	213.65	230.5
F1	9.16	125.12	210.95	230.7
Preferred	9.29	125.28	211.09	230.7
F2	9.29	125.75	211.36	229.4
StateSelected	9.29	125.19	211.15	230.5
NoOwn	67.50	506.87	739.45	289.0

Mean Poisson deviance at common eligible risk origins; smaller is better. Actual twenty-four-hour mean: 233.40 shares. Preferred selects F0/F1 using past validation; StateSelected permits F2; NoOwn removes own-history information. Paired calendar-week uncertainty excludes zero for the preferred one-hour improvement over F0, but includes zero at twelve and twenty-four hours.

D.1 Controls and contrasts

Age, spell age, publication-to-entry lag, and recent sharing enter as log-one-plus measures; trend compares the preceding hour with the hour before it. Additional controls are log topic attention, observation lag, historical tag intensity, current one-/six-hour tag activity, tag momentum/familiarity, New York hour interacted with weekend, section, tone, topic, and all eight incumbent attributes. Partisan and withholding controls are headline-only. The attention-filter uncertainty control is not relative-headline variance v. Apparent completion predicts continuation of the observation spell rather than economic survival.

With $\ell_n = \log(1 + \widehat{n})$ and $\ell_M = \log(1 + \widehat{M})$, the current-state specification is

$$g = \beta_1 \ell_n + \beta_2 \ell_n^2 + \beta_3 \ell_M + \sum_{p \in \{50,75,90\}} \gamma_p (\ell_M - c_p)_+ + \beta_4 \ell_n \ell_M. \tag{A16}$$

Knots c_p are full-risk outcome-independent quantiles. Date intercepts are profiled from events; eventless dates contribute no slope information. Non-event weights are one and offsets use actual eligible hours. Conditional quartiles use capacity deciles, topic, and age bands 0–1, 1–3, 3–6, 6–12, 12–24, and 24–48 hours. Standardization changes nonlinear terms consistently and averages over cells with at least thirty rows and nonzero support. Inference conditions on first-stage forecasts and scores.

D.2 Remaining-audience sensitivity

Table A6 compares horizons and the first-spell display measure under the unchanged preferred policy. The testing association is positive across total-audience horizons, but the six-hour

interval and display-count sensitivity permit either sign. Revision estimates are less precise. Display counts exclude later returns and cannot substitute for all future on-homepage exposure.

Table A6: RF1 audience horizons and measurement

Forecast	Action	Change (pp)	Article 95% CI	Article/week 95% CI
all, 24h	test	+0.0166	[0.0035, 0.0297]	[0.0030, 0.0302]
all, 24h	revision	-0.0023	[-0.0244, 0.0197]	[-0.0269, 0.0223]
all, 6h	test	+0.0132	[-0.0033, 0.0297]	[-0.0047, 0.0311]
all, 6h	revision	+0.0003	[-0.0274, 0.0280]	[-0.0271, 0.0277]
all, 12h	test	+0.0163	[0.0025, 0.0301]	[0.0019, 0.0307]
all, 12h	revision	+0.0102	[-0.0132, 0.0336]	[-0.0152, 0.0356]
home, 24h	test	+0.0072	[-0.0126, 0.0270]	[-0.0139, 0.0283]
home, 24h	revision	-0.0161	[-0.0402, 0.0079]	[-0.0412, 0.0090]

Net fifteen-minute probability contrasts in percentage points. "Home" is predicted sharing before the first apparent spell boundary. Other controls and actual-duration offsets follow the main specification.

E Static likelihood, measurement, and numerical estimation

E.1 Expanded cycling and linked outcomes

The strong first-event rule requires at least three transitions. On a binary sequence beginning with A, odd transition counts end at B and even counts end at A. Excluding two-transition reversions therefore selects against incumbent settlements. The expanded classification changes the terminal mark on 260 existing histories using exact event identifiers and attaches the corresponding challenger; it adds no independent article or duplicate initiation. The preceding exposure remains in the likelihood.

All added weak episodes are first observed interventions with clean recorded onset and no flagged wording ambiguity or recorded body update. Absence of an update flag is not verified content stability. In 150 of 260, the challenger is displayed for less than 1.5 minutes. The stricter rule also requires an onset bracket at most two minutes, no scrape gap above 150 seconds, and five to thirty minutes of cycling; the ten-minute variant raises the duration minimum. Neither rule verifies randomized assignment. The changing adoption shares show why uncertainty about detection must accompany cost interpretation.

Table A7: First binary cycling definitions

Definition	Tests	Completed	B adopted (%)
Strong first cycling	815	789	96.07
Expanded first cycling	1,075	1,042	72.74
Stricter expanded, five minutes	907	881	86.04
Stricter expanded, ten minutes	887	861	88.04

All definitions retain the underlying 36,234 histories, 1,380,928 intervals, and 340,846.76 eligible hours. Challenger adoptions number 758 in every row. Direct revisions remain 4,613; 149 multi-headline first events are separate binary-censoring endpoints. Stricter rows change event measurement and are not alternative fitted parameters for the main counterfactuals.

Of 1,075 expanded endpoints, 33 have unresolved outcomes. We sum over their possible winners rather than code them as keeping A. The 1,042 completed outcomes enter once, conditional on initiation. The broader 4,805 completed-episode sample has different observation, entry, and repetition rules and is excluded from this initiation likelihood.

E.2 Whole-history integration and normalization

For censored histories, equation (29) places the candidate integral outside the full survival product. For event histories, the revealed candidate density appears once and is multiplied by the survival of that same candidate and its terminal intensity. The candidate is conditional on the entry incumbent, not a claim that no editing occurred before entry. Initial-to-current incumbent offsets preserve its identity when measured wording changes slightly.

Conditional-independent truncated-normal coordinates restrict both feasible headline levels and dependence among attributes. Fixed difference units precondition numerical estimation; locations, common dispersion, action selection, and survival jointly determine the implied unselected distribution. A catalogue of observed pairs cannot replace that law without ignoring selection.

Fixed scrambled Sobol points from a bounded reference distribution and self-normalized importance weights integrate the history. The same draws enter every objective evaluation. The main fit uses 1,024 nodes, while the lower-resolution fits in Table A8 assess numerical sensitivity. They are separate estimates, not sources of mixed coefficients in the main parameter table. The final counterfactual snapshot has observed-history log likelihood $-19{,}457.594665$, reproduced at the same vector.

There are 21 free parameters: eight payoffs, eight candidate locations, common candidate

Table A8: Numerical refinement of the static specification

Draws	Cost	Prior SD	Prior intercept	Log likelihood
128	1.622309	0.0037071	0.0024155	-19,453.641
256	1.622126	0.0037068	0.0024102	-19,456.432
512	1.619930	0.0036521	0.0023669	-19,458.422
1,024	1.621307	0.0036985	0.0024116	-19,457.595

Nuisance parameters reestimated at each integration resolution, with the same economic restrictions. The 1,024-draw row supplies every primary structural and counterfactual result. Remaining integration variation is distinct from sampling uncertainty.

dispersion, prior standard deviation and intercept, review intensity, and testing cost. The common action-shock scale equals one. One forecast share defines one exposure-opportunity unit; direct-switching cost is zero; $\chi = 3$ is fixed; and the testing cost may equal zero. The loss weight in the general theory is not separately estimated. Article-sandwich covariance is based on complete-history scores and conditions on these restrictions and prepared inputs.

Table A9: Headline-payoff and candidate-distribution parameters

Attribute	$10^3\beta$	SE	Location a_j	Unit c_j
Signed sentiment	-0.3252	0.3658	-0.2789	0.3736
Democratic language	0.4439	0.6097	0.0319	0.2095
Information withholding	-0.8734	0.7783	-0.3297	0.2630
Readability	1.8583	1.2159	0.7869	0.1766
Specificity	-0.6830	2.2611	0.2430	0.1863
Personal consequences	0.6760	1.9009	-0.1106	0.1891
Agency	0.5402	0.5503	-0.2925	0.3142
Certainty	-0.1382	0.4559	1.2701	0.3533

Payoff coefficients and their conditional standard errors are multiplied by 1,000. Candidate locations are pre-truncation standardized coordinates, not realized mean differences. Their common dispersion and the remaining economic parameters appear in Table 2. Sentiment levels lie in $[-1, 1]$; all other levels lie in $[0, 1]$.

Recorded event times use operational display boundaries, and survival exposure is observed duration, not a count of deliberate keep choices. A constant-bin event contributes a continuous-time intensity density. Finite-interval predictions instead use $1-\exp[-(\ell_S+\ell_T)e]$ and allocate that probability by the competing intensities. The joint likelihood includes candidate densities and adoption, so it is not interchangeable with a discrete action-probability score.

E.3 Recovery and specification sensitivity

Recovery exercises simulate one persistent candidate, empirical audience paths and observation lengths, review arrivals, selected actions, and Gaussian learning. Two constant-review designs use five replications of 6,000 histories with positive or zero common prior intercept. Table A10 reports conditional testing-cost recovery; five replications are a screen rather than precise coverage evidence. Additional actual-scrape exercises show that detection can bias estimates even when the ideal terminal-mark likelihood is correctly implemented.

Table A10: Conditional cost recovery under the maintained static observation model

Generating prior	Cost bias	Cost RMSE	Profile coverage
Positive intercept	-0.018	0.088	5/5
Zero intercept	0.055	0.130	4/5

Each design has five replications of 6,000 histories and fixed precision. Cost coverage uses a nuisance-reoptimized profile at the generating value. These simulated economies verify conditional recoverability; they do not establish that weak observed cycles are randomized experiments.

The earlier strict, zero-intercept model estimated a cost near 1.751 but missed both strong-test adoption and the audience gradient. Expanding cycling alone reduced that cost to about 1.467. Adding a prior intercept with constant reviews gives the main specification. Allowing review arrivals to depend on age, clock time, and recent sharing improves some timing patterns, but alternative fits can drive prior uncertainty to its numerical floor and largely eliminate the audience channel. These specification comparisons use their own fitted vectors; they do not replace the common vector used for the November counterfactuals.

F Calibration and chronological assessment

In-sample predictions filter latent candidates using only preceding survival. Revealed candidates are used for conditional adoption evaluation, not inserted into predecision forecasts. Selected-candidate comparisons condition on observed action and event time. They therefore check the selected distribution rather than the unconditional proposal law.

Matching aggregate counts conceals timing errors. Table A12 groups actual eligible duration by time since entry. Tests and revisions are concentrated earlier than the constant-review model predicts. Counts beyond observed first events are not used in this diagnostic, so it does not validate the separate twenty-four-hour extension of the review clock.

Table A11: Selected headline differences: observed and model implied

	Revision		Testing	
Attribute	Observed	Fitted	Observed	Fitted
Signed sentiment	-0.0244	-0.0271	-0.0186	-0.0138
Democratic language	0.0043	0.0037	0.0079	-0.0026
Information withholding	-0.0211	-0.0190	-0.0189	-0.0236
Readability	0.0212	0.0223	0.0294	0.0240
Specificity	0.0296	0.0281	0.0324	0.0331
Personal consequences	0.0194	0.0252	0.0203	0.0220
Agency	0.0239	0.0187	0.0103	0.0296
Certainty	0.0292	0.0226	0.0325	0.0422

Mean challenger-minus-incumbent differences. Model means integrate candidate density, preceding survival, and observed action intensity, with adoption integrated out. This diagnostic uses the exact main parameter vector and is not a source of counterfactual candidates.

Table A12: Intervention timing on observed risk histories

	Tests		Revisions	
Elapsed hours	Observed	Fitted	Observed	Fitted
0-1	242	139.6	1,160	606.5
1-3	309	196.3	1,246	850.9
3-6	237	234.4	1,130	1,024.2
6-12	217	256.3	804	1,122.4
12-24	67	163.5	259	722.9
24+	3	61.4	14	296.3

Elapsed hours since first eligible origin. Predicted counts sum competing-event probabilities over the corresponding actual eligible durations.

Chronological training requires publication, all eligible follow-up, and recorded adoption completion before January 1, 2023; later publications form validation. The split contains 24,818 earlier and 11,416 later articles, with no overlapping histories requiring exclusion. The model is reestimated on the earlier group; Table A13 reports its later predictions without refitting. Its testing overprediction and adoption error are substantial.

Table A13: Later-period prediction after chronological training

Later-period moment	Observed	Predicted
Binary tests	162.0	369.9
Direct revisions	1,627.0	1,259.4
Challenger adoption (%)	56.41	82.77

11,416 articles and 408,093 intervals; 156 completed tests. The transported training fit is the same constant-review, common-intercept specification, not the full-sample vector reused as an out-of-sample estimate.

The weak-cycle share rises from 21.36 percent in training to 40.12 percent in validation,

changing observed adoption composition. Strong-test adoption remains high while expanded adoption falls. A common prior intercept cannot be presumed stable across that measurement change. Better full-sample adoption calibration under a richer review specification likewise does not ensure better later prediction.

The all-test audience comparison and the binary-only diagnostic use different informative dates and standardized rows. Their empirical point estimates are 0.015427 and 0.010516 percentage points. Article/calendar-week intervals are approximately [0.000805, 0.030048] and $[-0.003212, 0.024244]$. On identical binary cells, total intervention instead changes by -0.005781 points, with interval $[-0.034016, 0.022454]$; the test-share contrast is 0.828 points, with interval $[-1.124, 2.780]$. These are separate diagnostic regressions, so their coefficients need not form an accounting identity. The fitted contrasts in Table 3 hold the preceding survival filter and other current states fixed when changing remaining audience.

G Counterfactual simulation and exposure accounting

The common population contains 27,766 articles, or 76.63 percent of the estimation sample. Trajectories require modern forward-vintage forecast support and entry age at most twenty-four hours. The excluded 8,468 entries are older on average and have smaller entry audiences. Results for the common population are not extrapolated mechanically to them.

The reference path begins at first eligibility and reproduces the deployed all-channel forecast. At each later decision, original rolling thirty-minute and twenty-four-hour targets are recomputed using expected own counts from that path. Only the twelve quarter-hours preceding entry use actual counts. Peer states and vintage are frozen; age and clock time advance. Because the forecasts were trained under historical editorial behavior, treating this path as incumbent-headline opportunity is a maintained reference-demand assumption. The common twenty-four-hour outcome window and rolling twenty-four-hour decision horizon are different objects. Final headline outcomes include a late experiment's eventual settlement even when some settlement exposure lies beyond evaluation.

The main calculation uses 256 common persistent candidates per article and fifteen-minute bins. Competing first-event probabilities and Gaussian adoption are integrated analytically. With a constant within-bin hazard, expected event timing is integrated rather than assigned mechanically to the bin's start. A 512-candidate rule, a different seed, and a 7.5-minute grid

produce nearly unchanged marginal score contrasts. Six- and twelve-hour evaluation windows retain the rolling decision rule. Restricting reviews to original observed support is a separate sensitivity, not a claim that observation gaps are no-action decisions.

For a wording score g and menu $\mathcal{A}$, the exposure-weighted statistic is

$$\bar{g}^{\mathcal{A}} = \frac{\sum_i \mathbb{E}\left[\int_0^{24h} r_i^0(t) g(x_i^{\mathcal{A}}(t))\, dt\right]}{\sum_i \int_0^{24h} r_i^0(t)\, dt}. \tag{A17}$$

The fixed reference path is r_i^0; the expectation covers candidate, review, action, and signal realizations. During balanced testing, the score flow averages the separately transformed A and B scores. Final selection instead averages eventual selected scores equally across articles. The denominator never changes with the editorial menu.

With both options, the change in exposure negativity relative to keeping A separates into 0.004848 from direct revision, 0.000054 from temporary challenger exposure, and 0.001293 from post-test adoption. Revision alone contributes 0.006300. Adding testing reduces the direct-revision component; the experimental and post-test components almost offset that reduction. This accounting explains the small negative marginal exposure contrast without denying positive selection on latent quality.

Conditional structural bands use 128 antithetic draws from the matched article-sandwich covariance in stored parameter coordinates, restoring all fixed constraints. Each draw integrates 32 common candidates, and its deviation from the fitted coarse calculation is added to the 256-candidate point result. No draw violates the declared nonnegative-cost or numerical constraints. These bands hold forecast estimation, score measurement, detection, and the candidate family fixed.

An adoption sensitivity shifts the settlement index to match the observed completed-test mean while leaving fitted initiation unchanged. It is a settlement diagnostic, not a coherent reestimate. The testing-option exposure-negativity effect becomes approximately -0.000271, while exposure extremity becomes about 0.000005. Restricting review support reduces the added testing frequency from 5.865 to 2.500 percent. Primary results retain the estimated adoption rule and common clock extension.

G.1 Complete headline-composition outcomes

Tables A14 and A15 retain all eight measures and the two nonlinear transformations. The same initial headlines and candidate distribution enter every menu. Differences across sections arise from states under common parameters, not section-specific structural preferences.

Table A14: Final headline scores under each menu

Score	Keep	Revision	Testing	Both
Signed sentiment	-0.175582	-0.186045	-0.180003	-0.186717
Democratic language	0.495712	0.497805	0.496816	0.497962
Information withholding	0.422851	0.421506	0.420412	0.421268
Readability	0.754401	0.759644	0.758007	0.759901
Specificity	0.531674	0.539598	0.534631	0.539826
Personal consequences	0.377923	0.386239	0.381443	0.386746
Agency	0.405950	0.411737	0.409063	0.412607
Certainty	0.671682	0.680583	0.675367	0.681298
Negativity	0.256271	0.275929	0.263941	0.276752
Classifier extremity	0.363349	0.381895	0.370093	0.382571

Article-weighted expected final scores on the 27,766 common trajectories. Nonlinear scores are transformed separately on each headline before averaging.

Table A15: Exposure-weighted headline scores under each menu

Score	Keep	Revision	Testing	Both
Signed sentiment	-0.272547	-0.273732	-0.273343	-0.273888
Democratic language	0.531362	0.530698	0.531062	0.530934
Information withholding	0.406079	0.404426	0.404188	0.404148
Readability	0.758203	0.762270	0.761196	0.762383
Specificity	0.577275	0.580195	0.578575	0.580144
Personal consequences	0.430184	0.431812	0.431274	0.431869
Agency	0.538023	0.531473	0.534522	0.532560
Certainty	0.750074	0.747695	0.748479	0.748264
Negativity	0.318727	0.325026	0.321996	0.324921
Classifier extremity	0.397382	0.402614	0.399620	0.402713

Reference-audience weights over the common twenty-four-hour evaluation window. The reference denominator is identical across menus. Classifier extremity is $2|p_D - 0.5|$.

Gross payoff integrates comparative quality over challenger exposure. Direct revision uses the prior mean, temporary testing uses half the audience, and post-test settlement uses $\mathbb{E}[\Delta \mathbf{1}\{m > 0\}] = J(\mu, s)$. Test cost is k times initiation probability. Unobserved action utilities and payoff outside the common horizon are omitted from the systematic account. This convention prevents interpreting its net column as total welfare.

G.2 Section comparisons and reference scores

Table 5 uses all articles in U.S. news, World, and Opinion, reaggregating saved outcomes under the common fitted parameters. No section-specific model is estimated. Equal article weights define final scores; forecast audience weights define exposure scores, with a common denominator across menus within each section. The initial states, candidate distribution, and attention paths account for section differences. These are descriptive point comparisons without new subgroup uncertainty bands.

Table A16 reports the appropriate reference levels. Adding testing compares both options with revision alone; adding revision compares both options with testing alone. For each of the three scores, percentage changes equal one hundred times the score difference divided by that comparison-regime level, using unrounded values and matching final or exposure weights. Partisan-direction percentages are proportional changes in the classifier probability; its raw score differences remain the primary interpretation. The percentages do not measure changes in audience, engagement, political polarization, or ideological extremity.

Table A16: Reference headline scores for editorial-option comparisons

	Negativity	Democratic language	Readability
A. Revision-only comparison regime			
	Final / exposure-weighted score levels		
All articles	0.275929 / 0.325026	0.497805 / 0.530698	0.759644 / 0.762270
U.S. news	0.309992 / 0.305131	0.509230 / 0.577064	0.822812 / 0.823101
World	0.360633 / 0.381815	0.492468 / 0.502055	0.798272 / 0.770326
Opinion	0.353824 / 0.420012	0.491772 / 0.521041	0.637526 / 0.603111
B. Testing-only comparison regime			
	Final / exposure-weighted score levels		
All articles	0.263941 / 0.321996	0.496816 / 0.531062	0.758007 / 0.761196
U.S. news	0.298038 / 0.301109	0.508857 / 0.578286	0.827594 / 0.824212
World	0.354392 / 0.380309	0.490970 / 0.501802	0.800160 / 0.769868
Opinion	0.349454 / 0.420547	0.490368 / 0.521345	0.623161 / 0.595998

Each cell gives final / exposure-weighted score levels in the regime to which an option is added. Panel A supplies the denominator for adding testing; Panel B for adding revision. Populations and weights match Table 5. The numerical archive retains all unrounded regime levels and contrasts.

H Engagement-Responsive Homepage Placement

The main model holds the opportunity process common across headline actions, conditional on the reference state (Appendix A.4). A natural extension lets a better headline raise engagement, which can preserve homepage placement and attract additional future readers. Information then improves both response per reader and the audience path over which that response is earned. The publisher need not use the Twitter counts observed by the researcher when choosing placement.

Let X_t^r summarize reference story appeal and demand, and let $H_t \in \{0, 1\}$ indicate homepage display, with baseline headline exposure opportunities $r(X_t^r, H_t)$. Let q_A, q_B denote headline engagement rates. Define $\theta = q_B - q_A$ and net payoff $u_h = q_h - \lambda_L L_h$. A Gaussian prior and noisy experimental signal about θ use the updating equations in Section 3; known loss differences translate posterior engagement into net advantage. Beliefs about baseline engagement also enter the state.

Publisher-relevant engagement C_t and observed sharing Y_t satisfy a measurement bridge such as

$$\mathbb{E}[C_t \mid S_t, h] = r(X_t^r, H_t) q_h, \tag{A18}$$

$$\mathbb{E}[Y_t \mid C_t, \cdot] = \kappa_Y(X_t^r, H_t) C_t. \tag{A19}$$

Experimental engagement averages arm rates under balanced allocation. Current placement is predetermined; engagement then affects next-period placement:

$$\Pr(H_{t+1} = 1 \mid H_t = \ell, X_t^r, C_t) = \Lambda[\alpha_\ell + \beta_\ell' X_t^r + \gamma_\ell \log(1 + C_t)]. \tag{A20}$$

Here Λ is the logistic cumulative distribution function. Retention and return can have different responses. Setting both $\gamma_\ell = 0$ nests action-independent placement conditional on reference demand. Direct protection of active tests is an additional sensitivity, not required for engagement feedback.

Each headline has its own commitment value,

$$W_h(S_t) = \mathbb{E}\left[\sum_{j=0}^{T_i - t - 1} \delta^j r(X_{t+j}^r, H_{t+j}) u_h \mid S_t, \text{commit to } h\right]. \tag{A21}$$

The expectation integrates headline beliefs and the engagement/placement path induced by h. A higher engagement rate can increase both per-exposure payoff and later exposure, so $W_B - W_A$ need not equal a common audience times a posterior mean.

For a fixed-duration test with net payoff flow R^T, the action value is

$$Q_T^{\mathrm{P}}(S_t; k) = -k + \mathbb{E}\left[\sum_{j=0}^{L_T-1} \delta^j R_{t+j}^T + \delta^{L_T} \max_h W_h(S_{t+L_T}^+) \mid S_t, T\right]. \tag{A22}$$

Testing beats immediate commitment when $Q_T^{\mathrm{P}} > \max\{W_A, W_B\}$. After the experiment, adoption satisfies

$$W_B(S^+) > W_A(S^+). \tag{A23}$$

The posterior-mean rule is recovered when commitment differences are proportional to net advantage over a common exposure path. With common deterministic exposure, balanced testing, and no discounting, the shared baseline cancels to recover $MG - n|\mu|/2 - k$.

A sufficient state includes reference demand and beliefs, homepage status, current headline, persistent candidates, headline beliefs, and test phase. The action-dependent transition K^{P} integrates engagement, equation (A20), and private learning. For current net payoff $R(S, a)$, the recurrent problem is

$$V_t^{\mathrm{P}}(S) = \max_{a \in \mathcal{A}(S)} \left\{ R(S, a) - k\mathbf{1}\{a = \text{start test}\} + \delta \int V_{t+1}^{\mathrm{P}}(S') K^{\mathrm{P}}(dS' \mid S, a) \right\}. \tag{A24}$$

Active tests follow their protocol without repaying k each review. The single-intervention version commits after switching or test completion; a recurrent version preserves later choices and posterior beliefs. Flows plus continuation count each period's audience once.

Feedback does not uniformly favor experimentation. It can raise the payoff from discovering a better headline, favor immediate switching when the challenger already appears superior, or increase the allocation cost if experimental engagement weakens retention. Waiting can improve precision but lose placement under a weak incumbent. Both homepage states retain continuation: departure can hide choices without ending readership. Engagement-responsive observability alone does not change the common-audience comparison unless it alters relevant exposure, information, or options.

H.1 Future estimation with endogenous placement

The extension adds a publisher-engagement distribution and a placement response to the main likelihood. With structural parameters ζ, a complete-history factor is

$$\begin{aligned}\ell_t^{\mathrm{P}}(\zeta) = \pi_\zeta(a_t \mid S_t) f_C(C_t \mid S_t, a_t; \zeta) p_H(H_{t+1} \mid H_t, X_t^r, C_t; \zeta) \\ \times f_Y(Y_t \mid C_t, X_t^r, H_t, \mathcal{Z}_t; \zeta)^{C_t^Y} e_h(h_t^{\mathrm{obs}} \mid \mathcal{Z}_t; \zeta)^{O_t^h}. \end{aligned} \tag{A25}$$

Let K^r update reference demand and K^θ update private headline beliefs, candidates, and test phase. With status-observation kernel e_H,

$$\mathcal{L}_i^{\mathrm{P}}(\zeta) = \int \sum_{\text{hidden histories}} p_0(S_0; \zeta) \prod_t [\ell_t^{\mathrm{P}}(\zeta) K_\zeta^r K_\zeta^\theta e_H] \, d\mathcal{B}_i. \tag{A26}$$

The integral includes internal engagement, persistent private candidates, signals, and continuous beliefs. Signal updating appears once in K^θ. Missing headlines impose no observed wait decision, and incomplete tests receive no winner. Covered shares can inform demand while off-homepage histories evolve. Action-specific continuation uses equation (A24) and commitments W_A, W_B.

Identification must distinguish latent story appeal from headline-induced engagement: both can produce more sharing and longer display. Observed sharing and placement alone do not identify the feedback coefficient separately from appeal, measurement error, or private selection. Aggregate tweets cannot identify an unrestricted sharing-to-engagement bridge, placement-to-exposure mapping, and latent appeal simultaneously. Internal engagement, assignment records, or verified complete homepage snapshots would add discipline. The payoff/exposure normalizations and exact $k = 0$ benchmark remain necessary.

A staged estimator could fit reference demand, a restricted measurement bridge, and a parsimonious placement response before joint choice estimation. Each parameter vector requires solving the action-dependent dynamic problem; staging assumes policy-invariant first-stage primitives. Recovery exercises should apply the actual coverage mask and detector, jointly profile feedback, costs, and other parameters, and vary hidden editing and status error. No placement-feedback coefficient or counterfactual has been estimated.

Counterfactuals must regenerate engagement, placement, and exposure under each policy, replacing the main model's common audience path.