Pay Now or Wait to Unlock?
Cliffhangers and Monetization of Serialized Media

(Authors' names blinded for peer review)

April 2025

Abstract

Serialized media, including multi-episode books and shows, commonly uses a "wait-to-unlock" model wherein consumers can pay immediately for a new episode or wait for free access. We leverage a large-scale natural experiment on a U.S.-based fiction platform to examine the causal impact of different wait-time reductions across 10,000 series and over a million users. Employing a matching-based difference-in-differences framework, we show that shorter waits substantially increase aggregate consumption by both drawing in new readers *and* boosting engagement among existing ones. However, effects on paid unlocks vary with the extent of wait reduction: moderate cuts can drive revenue gains, while drastic cuts can cannibalize paid consumption—yet in certain series, deep reductions also intensify paid use among loyal readers. Crucially, we find that these heterogeneous effects are related to the "sequential complementarity" (operationalized as cliffhanger strength) across consecutive episodes, i.e., how strongly each episode's content drive immediate consumption of the next. We show that high complementarity mitigates cannibalization, leading consumers to purchase even when wait-times are drastically lowered. By highlighting how temporal "distance" between free and paid episodes interacts with sequential complementarity, our findings advance versioning theory and help platforms to devise book-specific wait to unlock strategies.

Keywords:
Serialized Media, Freemium, Temporal Versioning, Narratives

1 Introduction

Serialized media—such as books, TV shows, podcasts, or educational courses—are increasingly published as multi-episode works under a unified title. Digital media innovations and rising mobile consumption have amplified the popularity of serialized content, especially among the young,[1] with dedicated platforms emerging for comics (e.g., Webtoon), serialized fiction (e.g., Radish Fiction), and short-form videos (e.g., ReelShort) (MarketingCharts, 2019). Unlike standalone products, serialized media feature episodic storylines, recurring characters, and cliffhangers—narrative structures designed to encourage ongoing consumption by heightening anticipation for each subsequent installment (Kermode, 2000; Mittell, 2006; Linkis, 2021). These features of serialized media shape consumption patterns, as the perceived value of a future episode often depends on complementarities with respect to past episodes and how recently the previous episode has been consumed.

A prominent monetization strategy for digital serialized content is "wait-to-unlock" (or "wait-for-free") pricing, in which users may access the next episode either immediately for a fixed price or after a specified wait-time at a discounted price (often free). This approach differs from traditional temporal versioning, where distinct product versions (e.g., hardcover versus softcover books; movies in theaters versus DVDs or streaming) are sequentially released at different price points (e.g., Luan and Sudhir 2022). In the wait-for-free setting, the same episode effectively becomes a "premium" version if consumed immediately or a "freemium" version after the wait-time. As a result, wait-time becomes an important lever to modify consumers' costs of delayed consumption and the resultant monetization through a combination of free and paid consumption.

Despite the popularity of wait-to-unlock strategies on major platforms, research remains limited on how varying wait-times shape free and paid consumption in serialized media. Studies on versioning and freemium models typically treat quality differences as static (Shapiro

[1] Over two-thirds of 18 to 23-year-olds read episodic fiction, with 41% reading it every month according to research from the London Book Fair (Gynn, 2023).

and Varian, 1998; Varian, 2000), examining how free versions complement or cannibalize premium offerings (Lambrecht and Misra, 2017; Li et al., 2019). We extend this literature by highlighting a dynamic aspect of version "quality"—namely, how shortening the interval for free access can create non-monotonic effects on paid consumption. Unlike standard premium-versus-free versions, temporal versioning in serialized media presents ongoing choices, as each new episode can be unlocked after waiting or paying immediately.[2]

This paper addresses several novel research questions surrounding the wait-to-unlock paradigm. First, how does reducing wait-times affect overall consumption—does it primarily expand the breadth of readership by attracting new users at the extensive margin, or does it increase usage at the intensive margin (i.e., more episodes consumed within the user) among existing readers? Second, what is the impact on monetization and paid consumption—can shorter waits paradoxically boost total purchases within a consumer, or do they primarily cannibalize revenue by making it too easy to wait for free access? Finally, how do content attributes shape these outcomes—specifically, can text-driven "complementarity" between consecutive episodes moderate and explain the extent to which shorter wait-times either expand paid consumption or drive users towards wait-for-free rather than paid consumption?

To measure the role of wait-time in this dynamic setting, we leverage a large-scale natural experiment involving varying reduction in wait-times for select novels on a major U.S.-based serialized fiction platform. The platform implemented staggered, permanent wait-time changes for a subset of series: in some cases from 72 hours to 24 hours, and in others from 24 hours to 1 hour. Crucially, these adjustments occurred without announcement or concurrent promotional campaigns, limiting confounds such as marketing or seasonality. This setting allows us to isolate how consumption and purchase decisions respond to different intensities

[2]Choi et al. (2024) is a notable exception that studies wait-for-free as a promotion. They show that introducing wait-for-free promotion on serialized media effectively spurs both free and paid consumption by impacting the extensive margin for consumption. The wait-for-free draws new users to start reading more comics for free, but also increase overall revenue by inducing users, who are "hooked" to the book to buy paid episodes later. Our paper explores and generates novel insights around the intensive margins, demonstrates non-monotonic effects of wait-for-free, and through textual analysis provides greater insight into the mechanisms underlying the non-monotonic effects.

of wait-time reductions. We track user-level panel data over 15 months, covering more than a million users, 10,000-plus serialized works, and detailed records of whether each episode was unlocked by paying or waiting. Preliminary evidence suggests that while overall consumption tends to surge after a wait-time reduction, the effect on paid consumption varies markedly with the magnitude of the reduction.

Moreover, we leverage the raw textual data of the novels to understand the mechanism underlying consumer response to changes in wait times. We recognize that the aforementioned unique design of serialized media content—particularly the continuity or interconnectedness between consecutive episodes—can shape consumer responses to wait-times. Hence, we define *sequential complementarities* as the incremental value of the next episode that a consumer derives from consuming the current episode. Drawing on narratology theories, we hypothesize that sequential complementarities may push users to forgo waiting and pay for immediate access. To test this, we operationalize a novel measure of sequential complementarity based on "cliffhanger strength" using a large language model (LLM) that processes each episode's text. By quantifying the complementarity between consecutive episodes, we show how strongly (or weakly) an episode's content entices users to progress through the series without delay. We propose that such sequential complementarities (measured using episode level text data), will impact the trade-off between reaching a broader user base (through lower effective prices), while minimizing cannibalization of paid unlocks.

Identifying these effects requires navigating four major empirical challenges. First, to the extent that the platform selected certain series for wait-time reductions, there is potential selection bias if chosen series differ in unobserved ways from never-treated series. Second, the changes were staggered across time, creating an environment where standard difference-in-differences (DiD) estimators can yield biased estimates if treatment effects are heterogeneous or if control units themselves eventually receive treatment (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021). Third, the magnitude of the changes in wait times vary substantially across series; hence modeling the reduction as a binary "treat-

ment" indicator would obscure these differences and potentially confound small and large changes. Fourth, there is a need to capture episode-level textual content that underpins consumer urgency to consume content, which is not captured by conventional metadata (e.g., length or publication date).

We address these challenges using a matching-based panel DiD approach (Imai et al., 2023). Specifically, we match each treated series to control series with similar pre-treatment demand trajectories, age, and other covariates, ensuring balance on observed characteristics. We then implement a stacked DiD procedure centered on each series' treatment date, which mitigates confounds associated with staggered treatment adoption (Cengiz et al., 2019; Deshpande and Li, 2019; Deng et al., 2022; Guo and Liu, 2023). Further, the reduction in wait-time is modeled not as a binary indicator but as a percentage reduction, allowing us to estimate a non-monotonic relationship between the size of reduction and subsequent changes in both free and paid unlocks (Danaher et al., 2020; Zeng et al., 2022; Cook et al., 2023). Finally, to examine the role of sequential complementarities, we embed the large language model's episode-level text analysis into our empirical design, enabling us to tie purchase decisions directly to measured narrative continuity.

Our findings indicate that reducing wait-times can increase total consumption significantly, often through both an extensive margin (recruiting new readers) and an intensive margin (increasing engagement among existing users). However, the net impact on paid consumption is non-monotonic, depending on how much the wait-time is cut and on the content's underlying complementarities. Smaller wait-time reductions can draw more paying users without extensive cannibalization, while very large cuts sometimes undermine paid unlocks by making free access nearly frictionless—yet, interestingly, some very large cuts also spur deeper paid reading among certain loyal users. Analyzing textual complementarity shows that it shapes not only overall demand but also how severely wait-time reductions cannibalize purchases. In highly complementary series, even moderate waits can incentivize consumers to pay; by contrast, in low-complementarity series, reducing the delay can strongly

boost overall readership and sometimes paid consumption as well.

This paper makes three primary contributions. First, it extends the versioning and freemium literature by examining a distinctive form of temporal discrimination where identical episodes are priced differently based on wait-times. In contrast to traditional hardcover–paperback or multi-tier strategies, our results demonstrate that reducing the gap between premium and free versions can, under certain conditions, increase purchases (rather than uniformly cannibalize them) due to the inter-dependencies among sequential episodes. Second, we quantify textual complementarities using a large language model approach to operationalize narratological concepts. This methodology allows us to show how narrative structures affect user willingness to wait or pay. Third, our results offer managerial guidance on how platforms might tune wait-times to maximize engagement and revenue, emphasizing that content-rich, highly complementary series may benefit from moderate waits, whereas series with weaker continuity may need shorter delays to retain its audience.

The remainder of this paper is organized as follows. §2 positions our work in the context of research on serialized media, versioning and the economic impact of narrative features. §3 describes the institutional setting, nature of the wait-time reductions and our approach to measuring textual complementarity. §4 details the empirical strategy, including the matching-based DiD design, and reports our main findings. §5 presents the heterogeneous effects of different reduction magnitudes and the role of narrative continuity. Finally, §6 concludes with implications for platform strategy and limitations of our research.

2 Related Literature

This paper intersects with three main research streams: (i) consumption and monetization of serialized media, (ii) product versioning—particularly temporal versioning, and (iii) narratology. First, we build on the literature that investigates how multi-episode works (such as books, TV shows, or podcasts) engage audiences through episodic cliffhangers, charac-

ter continuity, and binge consumption patterns. Second, we draw upon versioning theories that examine how producers segment heterogeneous consumers by offering distinct access tiers—here, focusing on the dynamic nature of a "wait-for-free" window. Finally, we advance the narratology literature, which emphasizes how textual devices and narrative structure (e.g., suspense, emotional arcs) shape the user experience; we show that these elements can affect not just engagement depth but also consumers' decisions to pay for immediate access or wait for free. By integrating these three areas, our work highlights how content structure, timing of free availability, and consumer behavior jointly determine monetization outcomes in serialized media.

As discussed earlier, in contrast to standalone products, serialized content naturally segments the consumption experience into sequential installments that may be tightly linked by an overarching storyline or shared characters (Schweidel and Moe, 2016; Linkis, 2021; Zhao et al., 2022). These narrative structures, particularly when accompanied by cliffhangers or other suspense-based devices, can amplify user engagement and drive continued consumption. Research on binge-watching (Lu et al., 2019) underscores that the temporal spacing between episodes can critically influence purchase timing and loyalty. Several works study the phenomenon of binge consumption, exploring the implications on downstream behaviors such as responsiveness to advertisements, series completion and spillovers to other content on the platform (Schweidel and Moe, 2016; Lu et al., 2019, 2023; Godinho de Matos and Ferreira, 2020). Zhang et al. (2022) provides evidence of time-inconsistent preferences, where consumers intentionally choose to overpay for serialized content in order to curb future consumption (strategic self-control). Yet, while these works highlight the behavioral nuances of serialized consumption, they seldom isolate how varying the delay between free (or lower-priced) and paid access can shape monetization outcomes.

In parallel, the versioning (Shapiro and Varian, 1998; Varian, 2000) and freemium (Kumar, 2014) literatures offer frameworks for understanding how product differentiation—usually by quality or features—can segment a heterogeneous consumer base. Freemium studies have

emphasized the trade-off between consumer acquisition (through free usage) and potential cannibalization of premium sales (Kumar, 2014; Lambrecht and Misra, 2017; Li et al., 2019). Most such work, however, treats the "premium versus free" boundary as fixed, with little attention to dynamic or episodic settings. Temporal versioning of information goods has traditionally focused on release windows (e.g., hardcover vs. paperback books, theatrical vs. streaming films), but these windows are generally uniform across consumers rather than personalized to each individual's consumption clock (e.g., Luan and Sudhir 2022). Consequently, insights from classic release-window analysis may not fully capture the interplay of sequential content complementarities and customized wait-times.

In contrast to this literature, the wait-for-free approach in serialized media is an understudied form of temporal versioning wherein time itself acts as a lever for differential pricing. Although emerging empirical evidence points to the viability of such "wait-to-unlock" promotions on specific platforms (Choi et al., 2024), the broader question of how different magnitudes of wait-time changes shape total and paid consumption in serialized contexts remains open. Moreover, beyond basic measures of user behavior, the extent to which the underlying narrative content—for instance, strong cliffhangers or ongoing character arcs—might moderate the wait-time–consumption relationship is not well understood. Our study addresses these gaps by: (1) empirically examining how heterogeneous changes in wait-times can have non-monotonic effects on both paid and free engagement, and (2) quantifying textual complementarities across episodes to assess how content structure influences consumer decisions to purchase immediately or await free access. Thus, our contribution at the intersection of these two literatures—serialized-media consumption and dynamic temporal based product versioning—offers a novel lens on the monetization of serialized digital media.

Beyond the economic and managerial perspectives on serialized media, a parallel body of work in narratology explores how textual features drive audience engagement. Studies have shown that specific narrative devices—such as emotional arcs (Berger et al., 2021, 2023; Knight et al., 2024), surprise and suspense (Ely et al., 2015; Fong and Gui, 2024),

and narrative shapes (Toubia et al., 2021; Piper and Toubia, 2023)—significantly affect consumer engagement with narrative content. For serialized media, these elements that enhance sequential complementarity often manifest as cliffhangers: the deliberate suspension of a storyline at a suspenseful moment (Michlin, 2011; Schlütz, 2016; Linkis, 2021). Prior narratology research finds that cliffhangers can amplify emotional investment (Wirz et al., 2023), heighten suspense (Zillmann, 1995), and leave the narrative partially unresolved, thereby encouraging continued consumption (Poot, 2016). However, while such work highlights these textual devices, it typically stops short of linking them directly to economic outcomes—particularly whether and how cliffhangers can affect monetization.

Our paper extends this literature by rigorously quantifying cliffhanger "strength" and showing how that measure interacts with a monetary decision: whether readers pay for immediate access or wait for the free release of the next episode. Leveraging a large language model (LLM), we create a "cliffhanger score" that captures key narratological dimensions (e.g., suspense, surprise, emotional intensity, unresolved plot) from the preceding episode. This approach pushes beyond standard sentiment or topic analysis by focusing on sequential complementarities—the degree to which an episode propels the reader into the next installment. In other words, rather than viewing narratives purely as stylistic features, our method extracts continuity-driven intensity between episodes.

3 Institutional Setting and Data

The serialized fiction market consists of three key participants: authors, readers, and the two-sided platform. Independent authors publish series comprised of multiple episodes, which readers access through a mobile application. The market has seen rapid global growth, with notable platforms such as Wattpad and Kindle Vella.

We leverage data from a leading U.S.-based serial fiction platform specializing in the romance genre. The platform hosts over 10,000 series (i.e., books), each consisting of multiple

episodes and attracts over a million active users. The platform's user interface, which can be accessed only through a smartphone app, is illustrated in Web Appendix Section A.1. Revenue is primarily generated through users' episode purchases, with each series classified as one of three monetization categories: free, premium and wait-for-free (WFF). Free series allow immediate access to all episodes at no cost. Premium series follow a pay-per-episode model, where first several episodes are free, but subsequent episodes must be unlocked using an in-app currency ("Coins"). WFF series operate under temporal versioning. The key distinction between WFF and premium series is that WFF series allow users to unlock an episode for free once a pre-specified wait-time has elapsed *since the previous episode of the same series was unlocked*. Alternatively, users may bypass the wait-time and unlock episodes immediately by paying Coins, similar to premium series. Each episode costs 3 Coins (roughly 50 cents), and Coins can be purchased with real money.[3]

The wait-time for WFF series is set by the platform and varies across series, ranging from 1 to 72 hours. The same wait-time applies to all episodes and consumers of a given series, and the consumer must actively unlock the episode to "reset the clock" for the next free episode. For example, in a series with a 3-hour wait-time, a user may consume each episode in the series for free as long as she is willing to wait at least three hours *between each episode*. This prevents "wait-and-binge" behavior often observed in series with static release schedules, as a user returning after 12 hours will still have access to only one free episode rather than four. Moreover, there is no incentive for a user to purchase now to consume later (i.e., stockpile), as episodes eventually become free. To be clear, although firms have previously leveraged temporal versioning in contexts such as hardcover versus softcover books, the release timing in WFF is personalized, starting when a user consumes the previous episode and applies individually to each episode.

We leverage multiple datasets that cover user consumption, series and episode metadata,

[3]Users can also earn Coins through referrals, watching ads, or giveaway events. However, these alternative methods account for a negligible share of Coins used. Price per episode remains constant throughout our data.

as well as the text of each episode. The consumption panel data covers 15 months from October 1, 2020 to December 31, 2021 and details when and how (waited or purchased) the user consumed the episode. Series metadata include title, genre, author, sales type, date of first publication and the required wait-time. Episode metadata include sequence or position within the series, episode publication date and word count. The dataset also contains information on promotional activities where the platform offered coupons for specific series that can be used to unlock an episode, including the promotion dates and how many coupons were offered. Finally, we have access to the episode text, which we leverage to explore the mechanism driving our results.

3.1 Descriptive Statistics on Consumption and Payments

Figure 1 illustrates the distribution of series and consumption across the three sales types. Although WFF series constitute a third of all series on the platform, more than 85% of total and paid episode consumption in our dataset is generated by the WFF series.

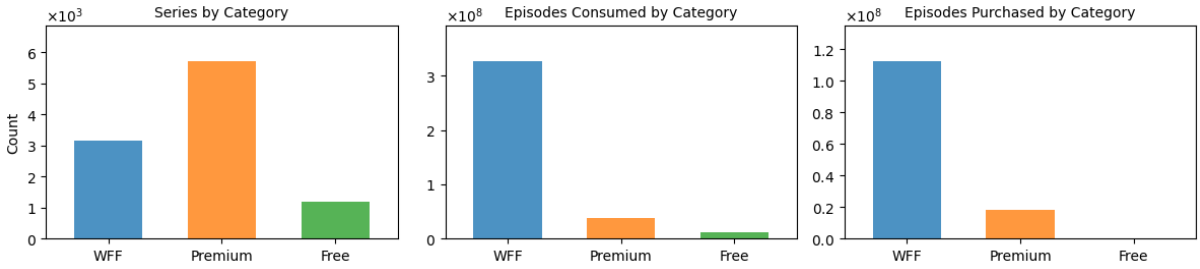


Figure 1: Distribution of series and episodes consumed across categories

Platform revenues are heavily concentrated among the most popular series. The left panel of Figure 2 illustrates the distribution of series by the size of their consumer base (log-transformed). The log-normal distribution reveals a pronounced concentration of readers on a small subset of series. The right panel of Figure 2 further highlights this concentration, showing episode purchase count from the ten most popular series, internally referred to as "mega-hits." These ten series, nine out of which are WFF series, account for approximately 40% of the platform's total revenue, despite representing only a small fraction of the series

catalog.

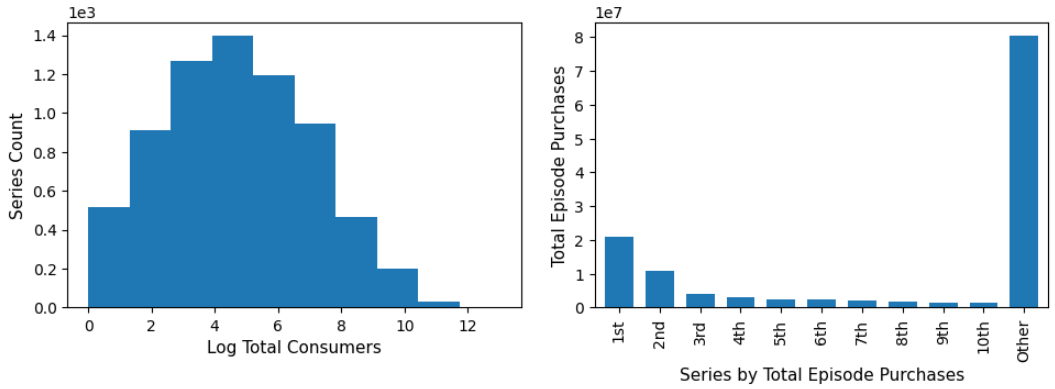


Figure 2: Distribution of readers across series (left) and revenues from the "mega-hit" series

Given the research goal of identifying the role of wait-times in consumption, we focus our analysis on the WFF series. Moreover, to reduce noise from tail end series that are rarely read, we filter for series with at least 1,000 episode accesses over the entire observation period. Our resulting dataset covers 1,822 WFF series and 308,681 users. The basic summary statistics are provided in Table 1. The median series contains 27 episodes, and the median user has read one series and 51 episodes during the observation period. She typically spends five days consuming a given series and reads one unique series on a consumption day. As is typical in media markets, the data is heavily skewed to the right, with a small number of heavy users driving the mean above the median, which we address in the analysis through log-transformation.

	Mean	**SD**	**25%**	**50%**	**75%**
Episodes per series	37.1	60.8	15.0	27.0	41.0
Series consumed per user	7.0	25.9	1.0	1.0	4.0
Episodes consumed per user	323.4	1078.9	13.0	51.0	186.0
Episodes purchased per user	96.7	326.2	1.0	9.0	53.0
Days spent per user per series	30.3	65.7	1.0	5.0	24.0
Series consumed per user per day	2.3	3.8	1.0	1.0	2.0

Table 1: Summary statistics on user consumption

3.2 Natural Experiment on Wait Times

Starting in November 2020, the platform began reducing wait-times for select series in a staggered manner. For example, users who had to wait 24 hours would now be able to access a free episode every hour after the reduction. Importantly, these changes were permanent, and both pre- and post-reduction wait-times varied across series, as detailed below.

According to the platform, they deliberately excluded its top ten "mega-hits," which generate the majority of revenues, from these changes. Instead, the reductions were targeted at older, longer, and less popular series that accounted for a smaller share of revenue. This strategy aimed to boost reader engagement by enabling quicker free consumption without jeopardizing significant revenue. Conditional on a series being selected for a reduction, the timing of the change and the post-reduction wait-times were determined randomly. Additionally, the platform made no prior announcements about the adjustments, ensuring that the changes were unanticipated from the readers' perspective. The firm also did not conduct marketing activities to specifically highlight the reduction following the change.

In our dataset, we identified 213 series that experienced wait-time reductions. We refer to them as *treated series*. we refer to the remaining 1,609 series, which did not experience changes to wait-times as *non-treated series*. Detailed breakdown of pre- and post-treatment wait-times is presented in Table 2, where the diagonal represents the number of non-treated series, and the off-diagonal represents the number of treated series. Notably, 160 out of the 213 treated series had their wait-times reduced to one hour.[4]

3.3 Quantifying Sequential Complementarities from Episode Text

Serialized media exhibit *sequential complementarities*, where consuming an episode enhances the value of the next due to the continuous narrative structure. For example, an episode that concludes with a suspenseful cliffhanger or an unexpected plot twist may compel the

[4]Among the 53 treated series with wait-times reduced to durations longer than one hour listed in Table 2, 29 were eventually reduced to one hour at a later date. Our analysis focuses on the first reduction each series received.

Pre/Post	1	2	3	4	5	6	7	8	10	12	24	36	48	72	All
1	1217	-	-	-	-	-	-	-	-	-	-	-	-	-	1217
2	1	6	-	-	-	-	-	-	-	-	-	-	-	-	7
3	14	-	17	-	-	-	-	-	-	-	-	-	-	-	31
4	6	-	-	30	-	-	-	-	-	-	-	-	-	-	36
5	1	-	-	-	4	-	-	-	-	-	-	-	-	-	5
6	6	-	-	1	-	5	-	-	-	-	-	-	-	-	12
7	1	-	-	-	-	-	1	-	-	-	-	-	-	-	2
8	1	-	-	-	-	-	-	5	-	-	-	-	-	-	6
10	1	-	-	-	-	-	-	-	2	-	-	-	-	-	3
12	46	-	3	2	-	-	-	-	-	139	-	-	-	-	190
24	66	-	23	1	-	-	-	2	-	8	163	-	-	-	263
36	-	-	-	-	-	-	-	-	-	-	-	1	-	-	1
48	16	-	1	7	-	-	-	-	-	2	3	-	17	-	46
72	1	-	-	-	-	-	-	-	-	-	-	-	-	2	3
All	1377	6	44	41	4	5	1	7	2	149	166	1	17	2	1822

Table 2: Number of series by wait-time (in hours) for pre- and post-reduction

consumer to immediately continue with the next episode. Importantly, sequential complementarity differs from the standard notion of complementarity, as the consumption *sequence* plays a critical role in the perceived value.

Consumption patterns in the data support the presence of sequential complementarities, i.e., users have a strong preference to consume content in their natural sequence. The probability of consuming episode e given that episode $e-1$ was consumed is 91%, whereas the probability of consuming episode e without having consumed episode $e-1$ is only 1%. Further, the probability of consuming episode $e-1$ after consuming episode e (in reverse order) is just 2%.[5]

As described earlier in the introduction and literature review, the narratology literature suggests that cliffhangers incorporate elements of suspense, surprise, emotional investment, and unresolved narrative arcs; hence for the purposes of our analysis, we use the *strength of the cliffhanger* as a single dimensional metric that captures multiple dimensions that factor into the strength of sequential complementarities across episodes. We quantify the cliffhanger strength by leveraging a large language model (LLM) to analyze the episode text.[6]

[5]Probabilities computed from a random sample of 5,000 consumers from the panel data.

[6]Research demonstrates that state-of-the-art LLMs outperform traditional machine learning models in psychological text analyses, such as detecting sentiment and emotions, even without fine-tuning (Krugmann

Specifically, we input the full text of episode $e-1$ into GPT-4o (version gpt-4o-2024-08-06) and ask for a single score ranging from 0 to 10 that reflects the strength of the cliffhanger leading into episode e. Details of the prompt are provided in Web Appendix Section A.2. The prompt details four criteria derived from the theoretical foundations discussed earlier, guiding the model's assessment. Importantly, our approach focuses exclusively on the content of the preceding episode $(e-1)$ to assess the decision to consume the next episode (e), consistent with the reality that consumers do not have access to the content of the next episode when making the consumption decision.

We compute scores for all episodes in the first season of series used in our main analysis, totaling 31,381 episodes.[7] The distribution and summary statistics of these scores are presented in Figure 3. Scores for individual episodes (left panel) range from 0.5 to 9, with a median of 6.5, indicating that most episodes feature strong cliffhangers. This aligns with the expectation that authors craft compelling cliffhangers to sustain audience engagement. As a sanity check, we separately visualize scores for the final episodes of each series. Since cliffhangers are intended to encourage continued consumption, we expect weaker cliffhangers in final episodes, where no subsequent installment follows. A two-sample t-test confirms this hypothesis ($p < 0.001$).

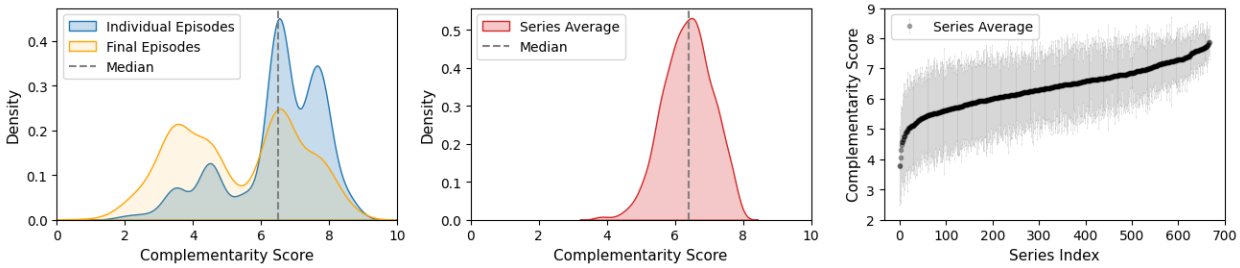


Figure 3: Complementarity score distribution for individual episodes and series-level average

At the series level (middle panel), average cliffhanger scores range from 3.8 to 7.9, with a median of 6.4. Finally, the right panel illustrates substantial within-series variation, where

and Hartmann, 2024; Rathje et al., 2024).

[7]Our dataset contains episode text from the first season; 63% of the series only have a single season. We obtain scores for all series that remain after propensity score matching, which is later described in detail in Section 4.

black dots represent series-level averages and gray bars denote standard deviations. To ensure the reliability of the LLM-generated scores, we conduct a validation study through an online survey on Prolific. The survey results confirm the robustness of the cliffhanger measures, as detailed in Web Appendix Section A.3.

4 The Effect of Wait-Time Reduction on Consumption

We begin with descriptive results on how consumption changes after the reduction in wait-times for the treated series. Thereafter, we discuss our empirical strategy involving a matching procedure and a stacked DiD approach for estimating the causal effect of wait-time reduction, accounting for the various empirical challenges in the data. After reporting the causal results based on our empirical strategy, we conduct a battery of robustness checks.

4.1 Changes in Aggregate Consumption with Wait-Time Reduction

Figure 4 compares the total number of episodes consumed (log-transformed) during the two-week period before and after the reduction for all treated series, irrespective of reduction magnitude. There is a significant increase in both paid and free consumption, suggesting that shorter wait-times may encourage greater consumer engagement and purchasing activity.

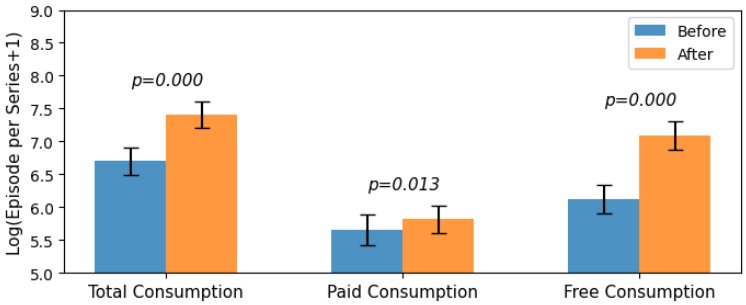


Figure 4: Comparison of total (paid + free), paid and free consumption (log-transformed) for treated series in a two-week window before and after the reduction

4.2 Empirical Strategy to Estimate the Causal Effects

While the above aggregate changes in consumption suggests that a reduction in wait-times led to higher free and paid consumption, this does not establish a causal relationship for several reasons. First, the selection of series for wait-time reductions could be endogenous, as the platform may have targeted series with specific characteristics or trends. Second, the analysis does not account for potential confounding factors, such as time-varying trends, concurrent promotional activities, or unobservable series attributes that could influence consumption. Third, this naive comparison assumes a uniform effect across all levels of treatment intensity, overlooking the possibility that the effect may vary depending on the magnitude of the reduction. To address these concerns, we present a robust causal model using a difference-in-differences framework and estimate the treatment effect controlling for a range of observable and unobservable factors.

To identify the causal relationship between wait-time and consumption, the analysis focuses on a 4-week window around the reduction.[8] By comparing the treated series to an appropriately constructed set of control series with no wait-time changes, we can estimate the effect using a difference-in-differences framework. But as mentioned earlier, our empirical context poses three challenges for conducting a DiD analysis. The first concerns selection into treatment. As the platform chose the series for wait time reduction, the treated and non-treated series may systematically differ, potentially leading to biased results if we naively compared the two groups (Rosenbaum, 2002). Importantly, we cannot even be certain of the direction of bias. For example, if the treated series inherently appeal less to consumers, then the estimated treatment effect would be downward biased, as the wait-time reduction has limited effect on getting users to consume the episodes. If the treated series contain more episodes, then the estimate would be upward biased.

The second challenge arises from the unbalanced panel data and variation in treatment

[8]The assumption is that any changes within this brief time period can be attributed to the wait-time reduction, controlling for a comprehensive set of features. We show robustness of our results to longer and shorter windows.

timing. Series are published and removed from the platform at different times, resulting in varying observation windows across series (only 6% of series are removed during the observation period). The unbalanced panel complicates the assessment of the parallel trends assumption, as missing observations can introduce differences in pre-treatment trends. Additionally, recent advances in econometric literature have shown that variation in treatment timing can bias average treatment effect (ATE) estimates in a two-way fixed effects (TWFE) model, especially when treatment effects are heterogeneous (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021; Borusyak et al., 2024). In such cases, the "forbidden comparison" of later-treated units to already-treated units may assign negative weights to certain sample treatment effects, causing the estimated ATE to diverge from the true effect.

The third challenge is that the pre-/post-reduction wait-times and hence the treatment intensity varies across the treated series. One may suspect that a reduction of large magnitude may have a different effect from a smaller reduction. There is also no reason to assume a linear relationship between the treatment effect and treatment intensity, as there may be potential non-linearities. Moreover, even if the absolute reduction magnitude is the same, a reduction from e.g., 12 to 11 hours may be different from a reduction from two to one hour. Therefore, a flexible model specification is required to estimate treatment effects as a function of treatment intensity while accounting for both pre- and post-reduction wait-times.

We address these challenges by using a panel-matching approach (Imai et al., 2023) and a stacked DiD model (Cengiz et al., 2019; Deshpande and Li, 2019; Baker et al., 2022; Butters et al., 2022; Deng et al., 2022; Guo and Liu, 2023) that estimates the treatment effect based on treatment intensity. We first match each treated series to a *matched control group* that consists of not-yet-treated series that are fully observed around the treatment timing and have similar propensity scores, or the probability of receiving treatment conditional on pre-treatment covariate histories.[9] Hence, treatment assignment is independent of poten-

[9]Consistent with the causal inference terminology, not-yet-treated series include the never treated as well as those series that, for a particular point in time, have not been treated yet (though they eventually become

tial outcomes conditional on observed covariates, approximating a randomized experimental design.

We then estimate the treatment effect using the stacked DiD method, which focuses on a fixed time window around the treatment event for each treated series, effectively creating a series of "mini" DiD analyses centered on the point of treatment adoption. The approach stacks these fixed time windows to form a consolidated dataset, within which the treatment effect is estimated using a DiD model that incorporates group specific fixed effects. By doing so, the stacked DiD model ensures that the estimation of treatment effects is grounded in a comparison of treated and control units within a narrowly defined temporal context, thereby restoring the validity of the parallel trends assumption and reducing the risk of biased estimates arising from heterogeneous treatment effects over time. This refinement allows for a more precise estimation of the treatment effect, accounting for the nuanced dynamics of staggered treatment adoption.

Finally, we estimate the treatment effect as a function of treatment intensity, defined as the difference between pre- and post-reduction wait-times (log-transformed) and include a higher order term to explore potential non-linearities. We also utilize a more flexible specification that estimates separate treatment effects for discrete levels based on treatment intensity. We provide additional details on how we address the empirical challenges in the following sections.

4.3 Constructing a Matched Control Group via Panel Matching

In order to address the potential systematic differences between the treated and the non-treated series, we create a control group for each treated series by matching it with not-yet-treated series that have similar probability of being treated. By making treatment independent of observed potential confounders, i.e. the conditional independence assumption: $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$, we can draw causal conclusions about the impact of reduced wait-

treated).

time by comparing the two groups. However, most existing applications with matching assume a cross sectional dataset using static features measured at a point in time (Hansen, 2004; Abadie and Imbens, 2011; Diamond and Sekhon, 2013). Applications using panel data typically compute the average of time-varying covariates over a static time-frame (Datta et al., 2018; Narang and Shankar, 2019; Deng et al., 2022) to fit the cross-sectional matching framework. However, this can miss out on important time-varying factors such as demand trends leading up to treatment that affect selection into treatment.

This is especially important in our setting, since some of the potential confounders are time-varying (e.g., number of episodes waited and purchased), and the variation in treatment timing makes it difficult to define a single pre-treatment period for the non-treated series. Moreover, matching on the average of time-varying covariates might match series whose covariates are similar on average, but exhibit very different temporal trajectories. For example, a series that is gaining traction among readers and one that is becoming less popular prior to treatment will clearly experience different effects from wait-time reduction.

Furthermore, since we have an unbalanced panel data with staggered treatment adoption, every matched series must be observed in the same time window. As an illustrative example, Figure 5 is a treatment variation heatmap, where each row represents a randomly sampled series and each column represents a week from our dataset. The red (blue) areas represent treated (non-treated) series-week observations, and white areas indicate no observation (i.e., series was not on the platform). We want to match each treated series to series that are fully observed and not treated (blue areas) around the treatment timing and are comparable in propensity scores. We therefore adapt propensity score matching for time-series cross-section data (panel-matching) developed in Imai et al. (2023).

Creating the Matched Control Group We now describe the matching procedure in detail. Let us denote a treated series s that receives treatment for the first time in period t as observation (s, t). For each treated observation (s, t), we identify a group of not-yet-treated

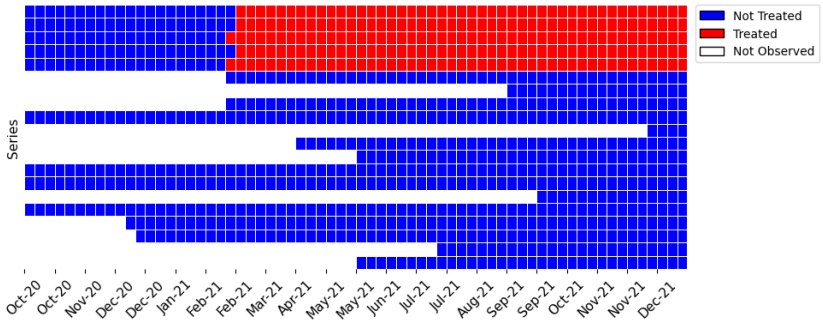


Figure 5: Treatment Variation Plot

units that are fully observed from time $t - L$ to $t - 1$. Figure 6 illustrates an example of how the groups are created when $L = 2$, indicated by the color of the boxes around the observations. In this example, treated series $s = 1$ is matched to series $s \in \{2, 3, 5\}$ over weeks $t \in \{1, 2\}$ (blue box). Note that the matched series are fully observed in the two weeks prior to the treatment timing of series $s = 1$. Series $s = 4$ is not included in the matched group because it is not observed in week $t = 1$. Similarly, series $s = 2$ is matched to series $s \in \{3, 4\}$ (red box). It is not matched to series $s = 1$ because it is already treated and $s = 5$ because it is not fully observed. In our case, we set $L = 4$, which assumes that adjusting for covariate trends up to previous four weeks removes the confounding. Formally, the matched control group for observation (s, t) is defined as

$$M_{st} = \{s' : s' \neq s,\ D_{s't'} = 0\ \forall\ t' = t, t-1, \ldots, t-L\} \tag{1}$$

where D_{st} is an indicator equal to 1 if series s is treated at time t (or earlier) and 0 if not.

Refining the Matched Control Group Next, we refine the groups based on propensity scores, the conditional probability of treatment assignment given observed pre-treatment covariates that can reasonably discriminate the treated and non-treated series (Rosenbaum and Rubin, 1983). The propensity score, e_{st}, is computed using a logistic regression:

Figure 6: Illustrative example of constructing the matched control group. The color of the boxes indicate the treated and matched control units included in the same group. For example, treated series $s = 1$ is matched to series $s \in \{2, 3, 5\}$ over weeks $t \in \{1, 2\}$ (blue box).

$$e_{st}(\{V_{s,t-l}\}_{l=1}^{L}) = Pr(D_{st} = 1|V_{s,t-1}, ..., V_{s,t-L}) = \frac{1}{1 + exp(-\sum_{l=1}^{L} \beta_l^{\intercal} V_{s,t-l})} \tag{2}$$

where $V_{s,t}$ is a matrix of observed static and time-varying covariates for series s in week t. These covariates include weekly count of waited and purchased episodes, promotional activities (defined as the number of coupons offered for the series), unique consumers, number of free, WFF and paid-only episodes in the series, age and required wait-time of the series. The use of endogenous pre-treatment variables (i.e., waited and purchased episodes) to compute propensity score is consistent with the existing research that utilize covariates such as lagged outcomes, consumer spending and income (Heckman et al., 1998; Dehejia and Wahba, 2002). These covariates serve as critical proxies for latent variables that might influence both the selection into treatment and the post-treatment outcomes of interest. By incorporating these variables, we aim to indirectly adjust for unobservable confounders. However, we cannot completely rule out the impact of such unobservables (Cinelli and Hazlett, 2020). In order to show that unobservables are not likely to be the driving factor of our results, we conduct a sensitivity analysis in Web Appendix Section A.7.4.

Given the fitted model, we compute the estimated propensity score $\hat{e}_{st}$ for all treated and their matched series. Among the series in the matched control group that belong to the same genre and whose propensity score distance to the treated unit is less than a defined caliper,

we select up to N series (or all units if fewer than N satisfy the criterion) with replacement.[10] Because the treatment timing varies across the treated series, potential concerns about over-reliance on specific control units from matching with replacement are mitigated.[11] Formally, the refined matched control group, M^*_{st}, for the treated observation (s,t) is given by

$$M^*_{st} = \{s' : s' \in M_{st},\ |\hat{e}_{st} - \hat{e}_{s't}| < C,\ |\hat{e}_{st} - \hat{e}_{s't}| \leq (|\hat{e}_{st} - \hat{e}_{s''t}|)^{(N)}\} \tag{3}$$

where $(|\hat{e}_{st} - \hat{e}_{s''t}|)^{(N)}$ is the N^{th} order statistic of the propensity score distance to the treated unit among the units in the original matched control group.

Covariate Balance Diagnostics Matching a treated unit to either a single or multiple control units is a common practice, with each approach presenting a tradeoff between bias and variance. One-to-one matching typically reduces bias by pairing each treated unit with its closest counterpart but increases variance due to the limited number of matches. In contrast, 1:N matching reduces variance by leveraging more data, but may introduce bias if the matched units are less similar to the treated unit or if certain controls are overused. To balance obtaining a sufficiently large sample against the risk of overfitting, we set $N = 10$.

The matching process yields eligible matches for 211 of the 213 treated series. Table 3 evaluates the balance of covariates and propensity scores before and after matching. The results show that the treated series and their matched control series are not significantly different in any of the variables based on p-values of the t-test. Figure 7 is a density plot of propensity scores before and after matching. Before matching, we see a greater density of control units with low probability of treatment as expected. After matching, treated and control groups are indistinguishable in terms of their treatment propensities, indicating a

[10] Since the majority (84%) of series are in the romance genre, we categorize series as romance or non-romance. Caliper is set to 0.5 of the standard deviation of the logit propensity score per guidance from Austin (2011). Top ten "mega-hit" series are excluded from matching as the platform deliberately did not alter their wait-times.

[11] As a precautionary measure, we allow each series to be used as a matched control series no more than ten times.

strong match.

	Treated	Control (Before Matching)		Control (After Matching)	
	Mean	Mean	p-value	Mean	p-value
Propensity Score	0.016	0.005	0.000	0.015	0.110
T1 Purchased Eps	2.912	3.048	0.290	2.774	0.254
T2 Purchased Eps	3.017	3.021	0.977	2.837	0.132
T3 Purchased Eps	3.113	3.015	0.448	2.953	0.178
T4 Purchased Eps	3.073	3.032	0.750	2.979	0.431
T1 Waited Eps	3.444	3.875	0.001	3.259	0.118
T2 Waited Eps	3.485	3.846	0.005	3.291	0.103
T3 Waited Eps	3.524	3.848	0.013	3.331	0.104
T4 Waited Eps	3.566	3.852	0.028	3.428	0.240
T1 Promotion	0.077	0.162	0.177	0.058	0.617
T2 Promotion	0.100	0.143	0.472	0.068	0.468
T3 Promotion	0.158	0.194	0.622	0.137	0.727
T4 Promotion	0.142	0.211	0.350	0.128	0.811
T1 Consumers	3.401	3.311	0.442	3.229	0.125
T2 Consumers	3.438	3.298	0.227	3.250	0.089
T3 Consumers	3.485	3.303	0.121	3.293	0.081
T4 Consumers	3.513	3.313	0.090	3.381	0.224
No. Free Eps	1.707	2.028	0.000	1.707	0.994
No. WFF Eps	3.642	3.450	0.002	3.635	0.918
No. Paid Eps	1.846	1.804	0.458	1.880	0.519
Series Age	6.147	5.000	0.000	6.171	0.714
Wait-time	21.536	7.257	0.000	20.435	0.250

Table 3: Covariate balance across treated and matched control series. T1 refers to trailing 1 week.

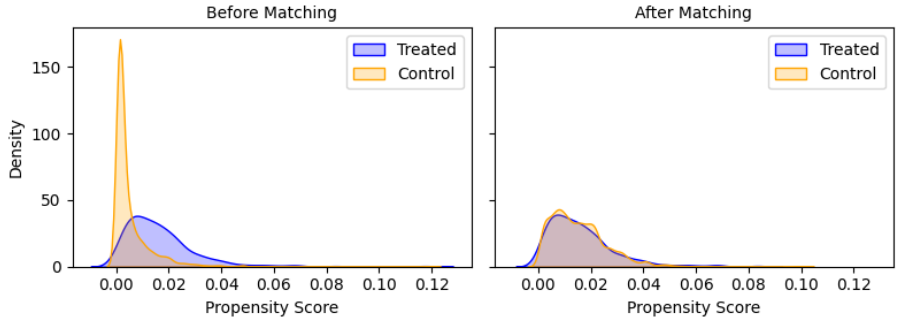


Figure 7: Propensity score distribution before and after matching

By matching series based on these detailed pre-treatment covariates, we argue that the matched series are balanced on the most critical factors influencing both treatment assignment and outcomes, leaving any remaining unobserved heterogeneity plausibly exogenous.

A potential confounding factor is series quality, which may influence both treatment assignment and outcomes. While series quality is inherently difficult to observe directly, we use trailing demand metrics, such as episodes purchased and waited, as proxies that reflect audience interest. Similarly, consumer engagement is captured through the number of unique readers and the mix of episode pricing categories (free, WFF, and paid-only), which account for heterogeneity in monetization strategies and consumption patterns. Series age accounts for lifecycle effects, such as differences in consumption patterns between newly launched and mature series, while genre captures topical differences across series. Finally, wait-time reflects the general willingness to pay for immediate access of the series audience, a key factor influencing both outcomes and treatment assignment. By ensuring balance across these detailed covariates, we mitigate systematic differences in unobserved factors that could bias our estimates, improving the plausibility of the unconfoundedness assumption.

Finally, the platform indicated that, conditional on being selected for a wait-time reduction, the post-reduction wait-times were randomly assigned without any systematic criteria. To empirically verify this, we regress the post-reduction wait-time on all covariates used to compute the propensity score, including time and genre fixed effects. The result in Web Appendix Section A.4 shows that none of the coefficients are statistically significant ($p > 0.1$) except for that on pre-treatment wait-time, verifying the randomness of post-reduction wait-time assignment conditional on treatment.

4.4 Difference-in-differences Model

To control for unobservable time-trends that may vary across units, we estimate the treatment effect in a DiD framework, incorporating fixed effects and covariates. However, in the presence of staggered treatment adoption and heterogeneity in treatment effects across units and time, the TWFE estimand can correspond to a non-convex weighted average of individual treatment effects, potentially leading to misleading conclusions. To address these issues, we employ a stacked DiD model, which has been widely applied in the marketing and

economics literature as a way to analyze data from a staggered treatment adoption design (Cengiz et al., 2019; Deshpande and Li, 2019; Baker et al., 2022; Butters et al., 2022; Deng et al., 2022; Guo and Liu, 2023).

We start by constructing event-specific datasets of equal length (4 weeks) around the reduction timing for each of the 211 treated series with eligible matches. The dataset includes the outcome and control variables of the treated series and its matched control series, which we denote as a *series group* (or cohort) consistent with the terminology from Deng et al. (2022). Note that although a control series may appear in multiple series groups, the corresponding data will vary depending on the reduction timing of respective series groups. We then stack these datasets together and estimate a DiD regression with group specific series and time fixed effects, which controls for self-selection on unobserved time-invariant factors. Gardner (2022) shows that this approach estimates a convex weighted average of the individual treatment effects.

To account for variation in reduction magnitude, we allow the treatment effect to vary based on treatment intensity. Extensive literature on prospect theory demonstrates that gains or losses in time are perceived in terms of proportional differences (Kahneman and Tversky, 1979; Leclerc et al., 1995; Abdellaoui and Kemel, 2014). For instance, Leclerc et al. (1995) show that individuals value a 15-minute reduction in wait-time from 1 hour significantly more than an equivalent reduction from 5 hours. Furthermore, research in psychophysics has shown that human perception of stimuli, including time, often follows a logarithmic relationship (Reichl et al., 2010; Haigh et al., 2021). Guided by these insights, we operationalize treatment intensity for series s in group g, R_{sg}, as:

$$R_{sg} = \log(\text{wait-time}_{sg,\text{pre}}) - \log(\text{wait-time}_{sg,\text{post}})$$

denoting the log-difference between wait-times or the log-ratio of wait-times before and after the reduction. This specification allows us to estimate the treatment effect as a function of

reduction magnitude while controlling for both pre- and post-reduction wait-times. Finally, this approach aligns with widely used methods in the economics literature, where treatment effects are modeled as a function of treatment intensity (e.g., Acemoglu and Finkelstein 2008; Danaher et al. 2020; Zeng et al. 2022; Cook et al. 2023).

Figure 8 displays the distribution of treatment intensity across the treated series in our data. The distribution is roughly normal, ranging from a minimum of 0.4 (33% reduction) to a maximum of 4.3 (99% reduction), with a median of 2.5 (92% reduction). The wide variation in treatment intensity highlights the platform's implementation of both modest and substantial wait-time reductions, creating a diverse setting to examine how varying levels of intensity influence consumption.

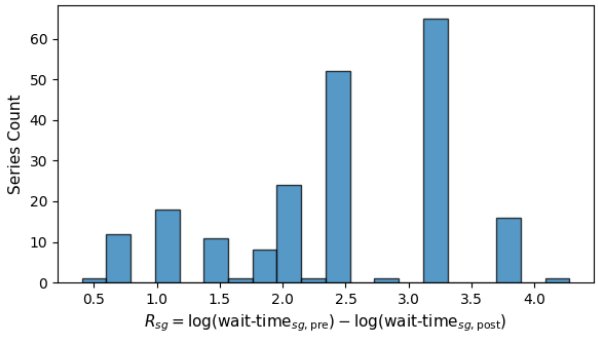


Figure 8: Distribution of treatment intensities

The stacked DiD model takes the following form:

$$\log(Y_{sgt}+1) = \beta_0 \cdot T_t \cdot \mathbb{1}(R_{sg}>0) + \sum_{n=1}^{2} \beta_n \cdot T_t \cdot (R_{sg})^n + \mathbf{X}'_{sgt} \cdot \gamma + \delta_{sg} + \nu_{gt} + \varepsilon_{sgt} \qquad (4)$$

The dependent variable, Y_{sgt}, captures different aspects of consumption activity for series s of series group g on day t, including the number of episodes consumed (total, paid and free) and the number of unique consumers engaging with the series. T_t is an indicator for the post-treatment period, and R_{sg} represents the treatment intensity for series s in group g. The primary coefficients of interest are β_0, β_1 and β_2. β_0 measures the baseline treatment effect for the treated series, while β_1 and β_2 capture the linear and non-linear marginal effect

of the reduction magnitude (R_{sg}) on the dependent variable. The overall treatment effect for a given R is computed as $\beta_0 + \beta_1 \cdot R + \beta_2 \cdot R^2$. The variance of the treatment effect is given by

$$\begin{aligned}\text{Var}(\text{TE}(R)) = \text{Var}(\beta_0) + R^2 \cdot \text{Var}(\beta_1) + R^4 \cdot \text{Var}(\beta_2) \\ + 2 \cdot R \cdot \text{Cov}(\beta_0, \beta_1) + 2 \cdot R^2 \cdot \text{Cov}(\beta_0, \beta_2) + 2 \cdot R^3 \cdot \text{Cov}(\beta_1, \beta_2)\end{aligned} \tag{5}$$

The standard error of the treatment effect is then $\text{SE}(\text{TE}(R)) = \sqrt{\text{Var}(\text{TE}(R))}$. For completeness, we include both results from specifications with and without the non-linear term.

The vector $\mathbf{X}_{sgt}$ contains observable controls, including the number of free, WFF and paid-only episodes, amount of promotional activity (measured as the number of episode coupons for series s offered on day t), as well as the sum of promotions over the trailing 7 days to control for potential lag effects. δ_{sg} represents group-specific series fixed effects (*Group-Series FE*), which account for time-invariant unobservable characteristics of series s in group g. ν_{gt} denotes group-specific time fixed effects (*Group-Time FE*), capturing unobservable time trends specific to group g on day t. By incorporating these fixed effects, the model effectively computes a DiD estimate for each group and applies variance-weighted aggregation to estimate the overall treatment effects (Baker et al., 2022). Finally, ε_{sgt} is the error term capturing all remaining influences.

We also estimate a more flexible version of the DiD model by estimating separate treatment effects for four discrete treatment intensity levels, denoted as $\mathcal{L}_n$, where $n \in \{1, 2, 3, 4\}$:

$$\log(Y_{sgt} + 1) = \sum_{n=1}^{4} \beta_n \cdot T_t \cdot \mathbb{1}(R_{sg} \in \mathcal{L}_n) + \mathbf{X}'_{sgt} \cdot \gamma + \delta_{sg} + \nu_{gt} + \varepsilon_{sgt} \tag{6}$$

where β_n captures the treatment effect for treatment intensities within level $\mathcal{L}_n$. The treatment intensities, which ranges from 0.4 to 4.3, are discretized into four approximately equal-width levels, as shown in Table 4. The last two columns of the table display the pre- and post-reduction wait-times of the treated series most frequently observed in each treatment

level. This model specification relaxes the strict parametric assumption between treatment intensity and the effect, allowing for greater flexibility in estimation. Hence forth, we refer to Equations 4 and 6 as parametric and semi-parametric specifications, respectively.

Treatment	No.	Treatment Intensity		% Reduction in Wait-time		Example Wait-time (hrs)	
Level	Series	Low	High	Low	High	Pre	Post
$\mathcal{L}_1$	31	0.4	1.3	33%	73%	3	1
$\mathcal{L}_2$	44	1.3	2.3	73%	90%	6	1
$\mathcal{L}_3$	119	2.3	3.3	90%	96%	24	1
$\mathcal{L}_4$	17	3.3	4.3	96%	99%	48	1

Table 4: Range of treatment intensity and percentage reduction for treatment levels

4.5 Results

The empirical analysis proceeds in three steps. We first examine the effect of wait-time reduction on consumption aggregated at the series level. Second, we analyze the extensive margin by measuring the effect on the number of users consuming any episode of the series. Third, we assess the intensive margin by measuring the effect on individual consumption of the existing consumers of the series. Finally, we conduct a battery of robustness checks to show validity of our results.

Series-level Aggregate Consumption

The net effect of wait-time reduction on *aggregate* consumption is theoretically ambiguous. Shorter wait-times can increase retention and total consumption, but they increase the risk of cannibalization, with paid consumption replaced by free consumption. Thus, the impact on paid consumption is an empirical question.

At a high level, we find that reducing wait-times increases aggregate consumption (total, paid and free) for the treated series, but the effects vary with treatment intensity. We analyze the change in total consumption, as well as paid and free consumption at the series-level over the two weeks before and after the wait-time reduction. Specifically, we estimate the stacked

DiD regression from Equations 4 and 6 using the daily panel data, where the dependent variables are the number of episodes consumed (total, paid and free) for a given series. We control for observable characteristics that may affect aggregate demand, as detailed in Section 4.4. Time-invariant series characteristics such as genre and age are absorbed by the group-series fixed effect.

The estimation results are presented in Table 5. Column (1) reports the results of a model incorporating linear effect of treatment intensity on total consumption. The positive and significant coefficients of $T \cdot \mathbb{1}(R > 0)$ and $T \cdot R$ indicate that shorter wait-times for free consumption lead to an increase in total consumption, proportional to the treatment intensity. Column (2) extends the analysis by incorporating the squared term of the treatment intensity to examine potential non-linear effects. While the coefficient of $T \cdot R^2$ is negative, suggesting that aggregate consumption increases with the treatment intensity at a diminishing rate, we must compute the standard error of the overall treatment effect using Equation 5. Upon doing so, we find that the treatment effect is significant at the 95% confidence level across the entire range of treatment intensity.

Column (3) presents a semi-parametric specification including a separate treatment effect for each treatment level as in Equation 6. To assess the magnitude of the treatment effect, let N_0 and N_1 represent the daily average consumption count before and after the reduction, respectively. Using the estimated level-specific treatment effect, we can compute $N_1 = e^{\hat{\beta}_n}(N_0+1)-1$, and the percentage change in the dependent variable as $(N_1-N_0)/N_0$. Given values of N_0 and $\hat{\beta}_n$, we assess the treatment effect for series with wait-times reduced to 1 hour from 3, 6, 24 and 48 hours (the most frequently observed series in each treatment level as noted in Table 4). For a series with 3-hour wait-time, if the daily episodes consumed before the reduction is at the mean ($N_0 = 272$), consumption would increase to 417, indicating a 53% increase. Similarly, a reduction to 1 hour would increase consumption by 76%, 109% and 99% for 6, 24 and 48 hour series, respectively.

Next, we estimate the impact of wait-time reduction on paid consumption, or the to-

tal number of episodes purchased for a given series. Increased total consumption would only be detrimental to revenues if it resulted in less paid consumption. While it is intuitive that shorter wait-times increase total consumption, their effect on paid consumption is ambiguous—determined by the balance between the base expansion effect and the negative cannibalization effect. The results in Columns (4)-(6) indicate that moderate wait-time reductions increase paid consumption. However, for large reductions (e.g., $R_{sg} \in \mathcal{L}_4$ such as a reduction from 48 to 1 hour), the estimated treatment effect turns negative, although it is not statistically significant. Coefficients from Column (6) suggest that reducing wait-times from 3, 6, 24, and 48 hours to 1 hour changes daily paid consumption by 44%, 24%, 16% and -17%, respectively.[12] This declining trend underscores an important boundary condition: excessively short wait-times may ultimately reduce aggregate paid consumption.

Finally, Columns (7)-(9) report the estimated effects on free consumption through waiting. As expected, we find that reductions in wait-times significantly increase free consumption, with the effect rising with the treatment intensity. Specifically, reducing wait-times from 3, 6, 24, and 48 hours to 1 hour increases daily free consumption by 56%, 99%, 152% and 174%, respectively. Figure 9 visualizes the treatment effects across the full range of treatment intensity, spanning from $R_{sg} = 0.4$ to $R_{sg} = 4.3$, demonstrating that the parametric and semi-parametric specifications produce qualitatively similar results.[13] To further validate these findings, we replicate the analysis at the episode level in Web Appendix Section A.6. The signs and magnitudes of the estimated treatment effects remain consistent with the series-level results.

Extensive Margins: Consumer Base or Breadth of Series

The prior aggregate analysis provides a measure of overall effects. However, we want to understand the mechanism underlying the effects. We therefore decompose the effect of

[12] The treatment effect for $\mathcal{L}_4$ is not significant at the 95% confidence level, possibly due to limited sample size. Episode-level analysis in Web Appendix Section A.6 shows a statistically significant negative effect.

[13] Parameters reported for the specification with best fit based on Akaike information criterion (AIC).

	Total Consumption			Paid Consumption			Free Consumption		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$T \cdot \mathbb{1}(R > 0)$	0.378***	0.201		0.484***	0.401		0.306**	0.174	
	(0.128)	(0.229)		(0.138)	(0.267)		(0.134)	(0.237)	
$T \cdot R$	0.110**	0.295		- 0.130**	-0.043		0.204***	0.342*	
	(0.047)	(0.195)		(0.052)	(0.237)		(0.050)	(0.203)	
$T \cdot R^2$		-0.041			-0.019			-0.031	
		(0.040)			(0.049)			(0.042)	
$T \cdot \mathbb{1}(R \in \mathcal{L}_1)$			0.424***			0.359***			0.440***
			(0.114)			(0.124)			(0.120)
$T \cdot \mathbb{1}(R \in \mathcal{L}_2)$			0.562***			0.210**			0.688***
			(0.094)			(0.099)			(0.098)
$T \cdot \mathbb{1}(R \in \mathcal{L}_3)$			0.734***			0.147**			0.918***
			(0.054)			(0.067)			(0.056)
$T \cdot \mathbb{1}(R \in \mathcal{L}_4)$			0.683***			-0.179			0.994***
			(0.111)			(0.119)			(0.130)
N Obs	64064	64064	64064	64064	64064	64064	64064	64064	64064
Adj. R^2	0.072	0.072	0.073	0.020	0.020	0.020	0.120	0.120	0.120
AIC	99886	99882	99872	155526	155528	155526	76692	76689	76694

Note: Robust standard errors clustered at the series level in parentheses. All regressions include observed control variables, group-series and group-time fixed effects. See Web Appendix for full estimation results.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 5: Treatment effect on series-level aggregate consumption

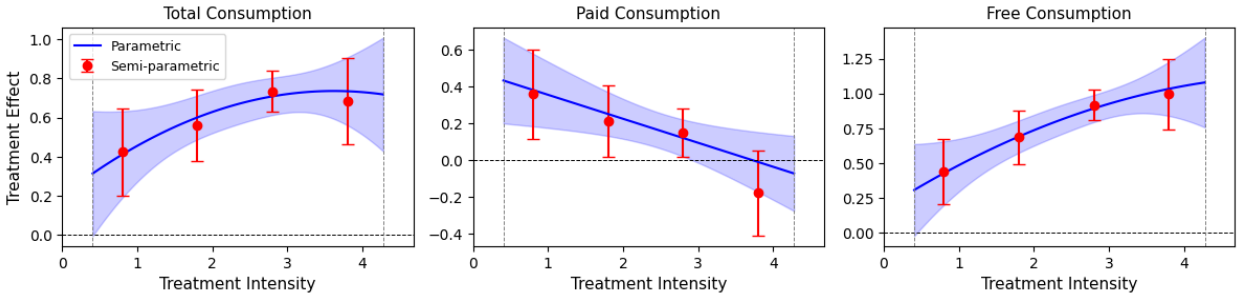


Figure 9: Treatment effect on series-level aggregate consumption by treatment intensity. Blue line and red dots represent estimates from the parametric (Eq. 4) and semi-parametric (Eq. 6) specifications, respectively. Shaded region and red error bars indicate 95% confidence intervals.

wait-time reduction on aggregate consumption into two components: the extensive margin (breadth of engagement) and the intensive margin (depth of engagement). This decomposi-

tion is crucial for understanding the drivers underlying the aggregate effects. The extensive margin captures changes in the size of the consumer base, reflecting how wait-time reductions influence new consumer acquisition and total engagement across all users. In contrast, the intensive margin isolates changes in the behavior of existing consumers, examining how reductions affect the number of episodes consumed and purchased by individuals already engaged with the series. By disentangling these two dimensions, we can identify whether the observed aggregate effects are driven by broader participation, deeper engagement, or both. This distinction is particularly important for informing platform strategy, as interventions that primarily expand the consumer base may require different marketing approaches than those that deepen engagement among existing users.

To analyze the extensive margin, we examine the effect on the number of users consuming any episode within a series. As before, we estimate Equations 4 and 6 using the four-week series panel data. The positive coefficient on $T \cdot \mathbb{1}(R > 0)$ and negative coefficient on $T \cdot R$ in Column (1) of Table 6 suggest an inverse relationship between the treatment intensity and effect: an increase in the consumer base for small reductions and a decrease for large reductions. Figure 10 confirms that results from the parametric and semi-parametric specifications are consistent. A possible explanation for this pattern is that small reductions broaden engagement breadth by attracting more new consumers, while larger reductions push consumers to expedite their consumption and exit the series, leading to a decrease in the consumer base despite the greater inflow. To quantify these effects, reducing wait-times from 3, 6, 24, and 48 hours to 1 hour changes the number of daily new consumers by 12%, 6%, 6% and -5%, respectively.

Intensive Margins: Depth of Series Consumption

Conditional on starting a series, how far does a consumer progress in the series and how many episodes does she purchase along the way? To evaluate the causal impact of wait-time reductions on within-individual consumption, we focus on consumers who consume episodes

	Total Consumers		
	(1)	(2)	(3)
$T \cdot \mathbb{1}(R > 0)$	0.311***	0.265	
	(0.113)	(0.208)	
$T \cdot R$	-0.118***	-0.070	
	(0.040)	(0.174)	
$T \cdot R^2$		-0.011	
		(0.035)	
$T \cdot \mathbb{1}(R \in \mathcal{L}_1)$			0.159
			(0.102)
$T \cdot \mathbb{1}(R \in \mathcal{L}_2)$			0.105
			(0.084)
$T \cdot \mathbb{1}(R \in \mathcal{L}_3)$			-0.020
			(0.045)
$T \cdot \mathbb{1}(R \in \mathcal{L}_4)$			-0.155**
			(0.077)
N Obs	64064	64064	64064
Adj. R^2	0.076	0.076	0.075
AIC	59594	59595	59627

Note: Robust standard errors clustered at the series level in parentheses. All regressions include observed control variables, group-series and group-time fixed effects. See Web Appendix for full estimation results.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 6: Treatment effect on series-level daily total consumers

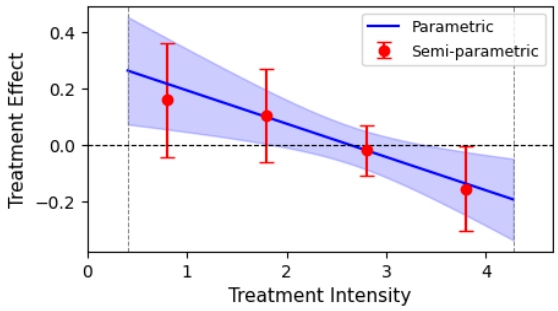


Figure 10: Treatment effect on series-level daily unique consumers by treatment intensity. Blue line and red dots represent estimates from the parametric (Eq. 4) and semi-parametric (Eq. 6) specifications, respectively. Shaded region and red error bars indicate 95% confidence intervals.

both within two weeks before and after the reduction for each treated and control series. We estimate Equations 4 and 6 using a two-period panel data with the number of episodes consumed per consumer (total, paid and free) as the dependent variable.

Columns (1)-(3) of Table 7 present the estimated treatment effects on total consumption per consumer. The results show that wait-time reductions significantly increase the number

of episodes consumed, with the effect growing monotonically with the reduction magnitude. This suggests that shorter wait-times encourage consumers to progress further in the series. Although there is a slight decline in the point estimate for $\mathcal{L}_2$ compared to $\mathcal{L}_1$, the difference is not statistically significant ($p = 0.27$). Level-specific estimates in Column (3) indicate that reducing wait-times to 1 hour increases total consumption per consumer by 24%, 7%, 64% and 135% for series with pre-reduction wait-times of 3, 6, 24, and 48 hours, respectively.

Columns (4)-(6) report the treatment effects on paid consumption per consumer. Note that from a versioning logic argument, paid consumption per consumer should *always* decline under reduced wait-times, as the primary effect would be cannibalization: consumers substituting from paid to free consumption. However, the results reveal a more nuanced pattern: paid consumption significantly increases for small and large reductions ($\mathcal{L}_1$ and $\mathcal{L}_4$) but decline for moderate reductions ($\mathcal{L}_2$). Specifically, reducing wait-times to 1 hour changes paid consumption per consumer by 31%, -20%, -1% and 68% for series with pre-reduction wait-times of 3, 6, 24, and 48 hours, respectively. This unexpected finding highlights the potential for shorter wait-times to drive additional monetization *within a consumer*, a result that has not been previously explored in the literature. We delve into the mechanisms driving this phenomenon in Section 5.

Finally, free consumption per consumer increases as wait-times shorten, with the effect rising monotonically with treatment intensity, consistent with expectations. Reducing wait-times to 1 hour increases free consumption by 11%, 18%, 89% and 147% for series with pre-reduction wait-times of 3, 6, 24, and 48 hours, respectively. Figure 11 plots the treatment effects across treatment intensity. Notably, the middle panel describes the U-shaped relationship for paid consumption.

4.6 Robustness Checks

We conduct a series of robustness checks to validate our empirical results. Here, we test for (i) whether the parallel trends assumption is violated and (ii) potential violations of the

	Total Consumption			Paid Consumption			Free Consumption		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$T \cdot \mathbb{1}(R > 0)$	-0.179	0.310		-0.166**	0.336***		-0.183	0.064	
	(0.132)	(0.194)		(0.078)	(0.109)		(0.158)	(0.240)	
$T \cdot R$	0.214***	-0.292*		0.066**	-0.452***		0.247***	-0.007	
	(0.044)	(0.156)		(0.028)	(0.090)		(0.053)	(0.188)	
$T \cdot R^2$		0.110***			0.113***			0.056	
		(0.032)			(0.019)			(0.038)	
$T \cdot \mathbb{1}(R \in \mathcal{L}_1)$			0.202**			0.162***			0.090
			(0.082)			(0.059)			(0.130)
$T \cdot \mathbb{1}(R \in \mathcal{L}_2)$			0.063			-0.157***			0.148
			(0.097)			(0.046)			(0.117)
$T \cdot \mathbb{1}(R \in \mathcal{L}_3)$			0.448***			-0.006			0.566***
			(0.043)			(0.024)			(0.049)
$T \cdot \mathbb{1}(R \in \mathcal{L}_4)$			0.762***			0.304***			0.786***
			(0.073)			(0.050)			(0.086)
N Obs	740722	740722	740722	740722	740722	740722	740722	740722	740722
Adj. R^2	0.006	0.006	0.006	0.001	0.001	0.001	0.008	0.008	0.008
AIC	1800840	1800726	1800751	1981589	1981495	1981423	1919098	1919075	1919038

Note: Robust standard errors clustered at the series level in parentheses. All regressions include observed control variables, group-series and group-time fixed effects. See Web Appendix for full estimation results.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 7: Treatment effect on consumption per consumer

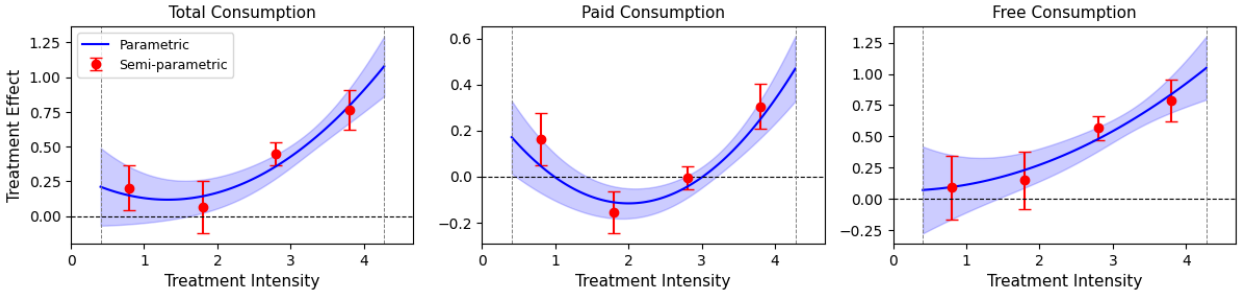


Figure 11: Treatment effect on consumption per consumer by treatment intensity. Blue line and red dots represent estimates from the parametric (Eq. 4) and semi-parametric (Eq. 6) specifications, respectively. Shaded region and red error bars indicate 95% confidence intervals.

stable unit treatment value assumption (SUTVA). In addition, we conduct a number of robustness checks including alternative time windows around treatment, pseudo-treatment ef-

fects, subsample analyses, sensitivity to unobserved confounders, and a fully non-parametric estimation of treatment effects.

Parallel Trends

Identification of the causal effect of wait-time reduction on consumption relies on the parallel trends and no anticipation assumptions. The parallel trends assumption requires that, in the absence of treatment, the treatment and control groups would have followed similar trends. The no anticipation assumption mandates that outcomes were unaffected by the impending treatment prior to its implementation. If these conditions hold, time-varying unobservables are captured by the trends in the control group, yielding unbiased estimates. In our setting, the platform reduced wait-times for selected series at different times without prior notification, preventing users from altering their consumption or purchasing behaviors strategically.

To formally test the parallel trends assumption, we follow a standard approach in the literature that leverages pre-treatment panel data (Angrist and Krueger, 1999; Bronnenberg et al., 2020). Specifically, we use the two-week pre-treatment period to estimate

$$\log(Y_{sgt} + 1) = \alpha_0 \cdot t + \alpha_1 \cdot t \cdot R_{sg} + \mathbf{X}'_{sgt} \cdot \gamma + \delta_{sg} + \varepsilon_{sgt} \tag{7}$$

where the dependent variable is series-level aggregate daily consumption (log-transformed). Here, α_0 captures the common trend, and α_1 measures deviations from this trend moderated by treatment intensity during the pre-treatment period. This tests whether treatment intensity moderates the pre-treatment trends, a generalized form of the parallel trends assumption. As shown in Table 8, deviations for treated series (α_1) are not statistically significant. Thus, we fail to reject the null hypothesis that treated and control series exhibit similar pre-treatment trends, supporting the parallel trends assumption. The results are robust to a longer four-week pre-treatment window.

Number of Episodes	Total Consumption	Paid Consumption	Free Consumption
Trend (α_0)	0.008***	0.002	0.011***
	(0.003)	(0.004)	(0.003)
Trend · R (α_1)	-0.002	-0.005	-0.001
	(0.002)	(0.003)	(0.002)
No. Free Eps	-0.489	-0.232	-0.509*
	(0.303)	(0.452)	(0.306)
No. WFF Eps	0.706**	0.697	0.635***
	(0.303)	(0.500)	(0.240)
No. Paid Eps	0.054	-0.004	0.067
	(0.126)	(0.144)	(0.088)
Promotion	0.116***	0.138***	0.102***
	(0.022)	(0.024)	(0.022)
T7 Promotion	0.160***	0.149***	0.164***
	(0.035)	(0.028)	(0.038)
N Obs	32032	32032	32032
Adj. R^2	0.063	0.024	0.104

Note: Robust standard errors clustered at the series level in parentheses. All regressions include group-series fixed effects.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 8: Parallel trends

Stable Unit Treatment Value Assumption (SUTVA)

For the causal interpretation of treatment effect estimates in a DiD analysis, the SUTVA assumption that potential outcomes of each unit must not be affected by the treatment assignment or outcomes of other units must be satisfied (Rosenbaum and Rubin, 1983). In our context, SUTVA could be violated if consumers with limited time and resources shift their consumption from control series to treated series (substitution), or conversely, if increased engagement with treated series enhances consumption of control series (complementary).

To assess potential violations of SUTVA, we demonstrate that (1) the results remain robust when restricting the control group to series with minimal overlap in their consumer base with treated series, (2) the results are consistent even when explicitly controlling for the overlap, and (3) individual consumption patterns for control series among consumers of treated series do not significantly differ before and after treatment. The assumption underlying the first two approaches is that spillover effects, if present, are more likely for control series that have a higher proportion of consumers shared with the treated series.

First, we estimate Equation 6 using series-level aggregate consumption as dependent variables, restricting the control group to non-treated series with an overlap of less than 10%. The overlap measure, $overlap_{sg}$, is defined as the proportion of consumers of non-treated series s in group g who also consumed an episode of a treated series in the same group during the two weeks prior to treatment ($overlap_{sg} = 0$ for treated series). The results are presented in the first three columns of Table 9. Even restricting the control group to non-treated series with minimal potential for spillover effects, the estimated treatment effects are similar and qualitatively consistent as in the main specification in Table 5.

Second, we add an interaction term, $T_t \cdot overlap_{sg}$, to Equation 6 to explicitly account for potential spillover effects. The coefficient on this term quantifies spillovers, isolating the treatment effect. This approach of addressing interference has been widely adopted in the economics and marketing literature (e.g., Clarke 2017; Jo et al. 2020). The results, shown in the last three columns of Table 9, remain consistent with the main findings. When free consumption is the dependent variable, we also find that the effect of reduction in the control series is positive and statistically significant. By explicitly controlling for spillover effects, this approach confirms that the main results are robust to potential violations of SUTVA.

	(1) Subsample Analysis			(2) Controlling for Overlap		
DV: Consumption	Total	Paid	Free	Total	Paid	Free
$T \cdot \mathbb{1}(R \in \mathcal{L}\_1)$	0.457***	0.361***	0.477***	0.426***	0.360***	0.442***
	(0.111)	(0.124)	(0.112)	(0.115)	(0.124)	(0.119)
$T \cdot \mathbb{1}(R \in \mathcal{L}\_2)$	0.597***	0.287***	0.718***	0.566***	0.213**	0.693***
	(0.092)	(0.095)	(0.097)	(0.093)	(0.099)	(0.097)
$T \cdot \mathbb{1}(R \in \mathcal{L}\_3)$	0.739***	0.152**	0.924***	0.730***	0.145**	0.914***
	(0.054)	(0.067)	(0.056)	(0.054)	(0.067)	(0.056)
$T \cdot \mathbb{1}(R \in \mathcal{L}\_4)$	0.677***	-0.184	0.989***	0.677***	-0.183	0.986***
	(0.113)	(0.119)	(0.131)	(0.112)	(0.119)	(0.131)
$T \cdot overlap$				0.039**	0.024	0.045***
				(0.015)	(0.018)	(0.015)
N Obs	58716	58716	58716	64064	64064	64064
Adj. R^2	0.074	0.020	0.124	0.073	0.020	0.121

Note: Robust standard errors clustered at the series level in parentheses. All regressions include observed control variables, group-series and group-time fixed effects.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 9: Check for potential violation of SUTVA using consumer base overlap

As a final piece of evidence, we examine changes in individual-level consumption to provide direct support for SUTVA. Specifically, we assess whether consumers of a treated series engage with the same number of series and consume (total, paid, and free) a similar number of episodes from the matched control series before and after treatment. Table 10 shows that consumption patterns remain largely consistent across periods. While consumers engage with slightly fewer series post-treatment, the number of episodes purchased or waited for within a series shows no significant differences, as indicated by paired t-test p-values.

	Mean (Before)	Mean (After)	p-value
No. Series	1.223	1.105	0.000
No. Eps Read	9.574	9.232	0.375
No. Eps Purchased	0.912	0.959	0.689
No. Eps Waited	8.662	8.273	0.286

Table 10: Individual consumption patterns of non-treated series before and after reduction

Additional Sensitivity Checks

We perform a series of additional analyses to verify the robustness of our main findings. While the full details appear in Web Appendix Section A.7, we summarize the analyses and results here:

i. Alternative time windows. We test whether the results depend on the choice of a four-week window around the treatment. Re-estimating treatment effects with both shorter (two-week) and longer (eight-week) windows yields qualitatively similar estimates, suggesting that our conclusions are robust to the definitions of pre- and post-treatment periods.

ii. Pseudo-treatments. Next, we conduct two falsification tests: (1) randomly assigning the treatment indicator to control series (thereby excluding the truly treated ones), and (2) shifting the actual treatment date earlier so that the real policy change falls outside the

examined window. Both tests produce no significant effects, ruling out the possibility that our DiD framework merely captures spurious correlations or arbitrary model artifacts.

iii. Subsample analyses. To address potential confounds from promotional activities and variation in post-reduction wait-times, we replicate the analysis under two restricted samples: (1) series with no concurrent promotions and (2) treated series that share a uniform post-reduction wait-time of one hour. In both cases, the estimated treatment effects remain consistent with our main results, indicating that our findings do not hinge on promotional timing or heterogeneous final wait-times.

iv. Sensitivity to unobserved confounders. We apply the sensitivity analysis of Cinelli and Hazlett (2020), which quantifies how strong an unobserved factor would need to be to invalidate our treatment effect estimates. Even under extreme assumptions, the predicted bias is insufficient to overturn the results, reinforcing that unobserved confounding is unlikely to drive our conclusions.

v. Non-parametric estimation. Lastly, following Callaway et al. (2024), we estimate treatment effects without imposing specific functional forms on treatment intensity. The non-parametric estimates closely mirror the parametric and semi-parametric patterns, offering additional support that our main findings are not artifacts of model specification choices.

Collectively, these checks confirm that our key takeaway—shorter wait-times drive significantly higher total consumption and exhibit non-monotonic effects on paid consumption—remain robust to a range assumptions and analyses.

5 The Role of Sequential Complementarities

Having established the baseline causal effects, we now investigate why wait-time reductions impact paid consumption unevenly across series and episodes. Specifically, we incorporate

the sequential complementarity metrics derived from the episode level textual data developed in Section 3.3 to assess how the narrative tie from one episode to the next shapes user decisions to pay versus wait.

5.1 Sequential Complementarities and Episode-level Aggregate Consumption

To examine how sequential complementarities influence consumer response to wait-time reductions, we estimate treatment effects on episode-level aggregate consumption, separating episodes into high- and low-complementarity groups using a median split based on their complementarity scores. As described in Section 3.3, these scores reflect how strongly an episode's text "hooks" the reader into the following episode. Figure 12 displays the treatment effects on aggregate consumption (paid vs. free), estimated using Equation 4. Full results are in Web Appendix Section A.8.

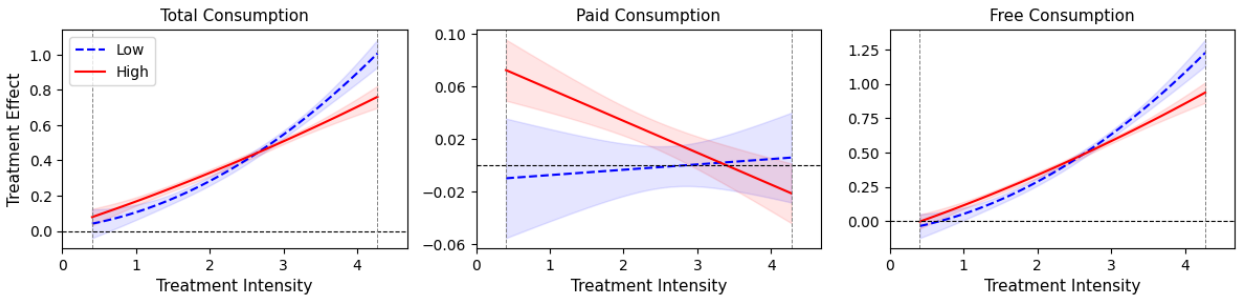


Figure 12: Heterogeneous treatment effect on episode-level aggregate consumption. Blue dotted and red solid lines represent estimates from the parametric specification (Eq. 4) for episodes below and above median complementarity, respectively. Shaded regions indicate 95% confidence intervals.

The key takeaways are: (1) High-complementarity episodes exhibit a significant rise in paid consumption following a wait-time reduction, particularly at moderate cut levels. However, as treatment intensity becomes very large, the impact on paid consumption diminishes. (ii) Low-complementarity episodes show no significant change in paid consumption overall, suggesting that weaker narrative ties do little to spur immediate unlocks when wait-times are

shortened; (iii) Free consumption increases for both high- and low-complementarity episodes, with the effect growing monotonically as the wait-time reduction becomes very large. Overall, these patterns imply that episodes with stronger continuity benefit more from moderate cuts in wait-times, whereas extremely large cuts begin to reduce the impetus for paying, as free access becomes too convenient.

5.2 Sequential Complementarities and Intensive Margins

Next, we analyze the intensive margin—how consumption changes within a consumer—by classifying entire series into low- and high-complementarity categories, based on a median split of each series' average complementarity score. Figure 13 plots the treatment effect on per-consumer consumption (paid and free) across varying wait-time intensities, again using Equation 4. Full results are in Web Appendix Section A.8.

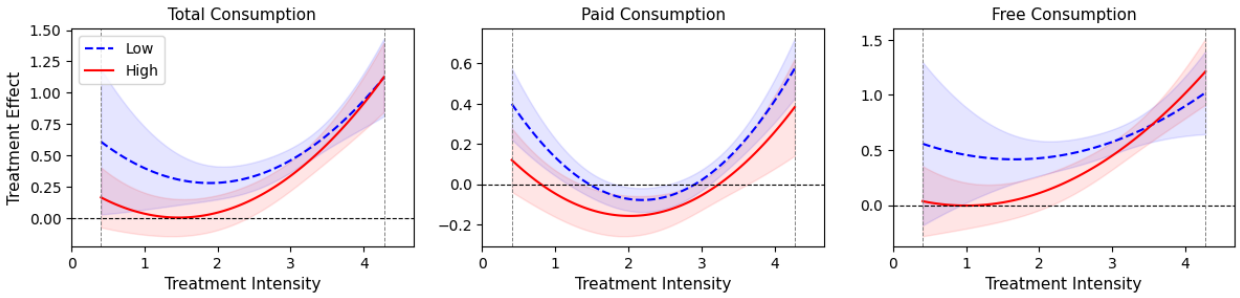


Figure 13: Heterogeneous treatment effect on consumption per consumer. Blue dotted and red solid lines represent estimates from the parametric specification (Eq. 4) for series below and above median complementarity, respectively. Shaded regions indicate 95% confidence intervals.

The key takeaways are: (1) Low-complementarity series show a larger jump in consumption (both paid and free) after wait-time reductions than high-complementarity series, suggesting that shorter waits help "rescue" users from low-value episodes and retain them longer—thereby creating more downstream opportunities to monetize. (ii) High-complementarity series are more capable of holding user interest even with longer waits, so the incremental effect of shorter wait-times on within-consumer consumption is more muted.

Overall, the results suggest that while high-complementarity episodes can elicit immediate purchases in a single instance (Section 5.1), entire low-complementarity series gain more total consumption–especially on the intensive margin–when wait-times are shorter, enabling users to bypass potential "drop-off points" quickly.

5.3 Sequential Complementarities and Individual Consumption Decisions

Finally, to investigate how wait-times influence individual consumption decisions and how sequential complementarities moderate this relationship, we estimate the following fixed-effects logistic regression model:

$$\text{Logit}\Big[P(Y_{iest}=1)\Big] = \beta_0 + \beta_1 \cdot W_{st} + \beta_2 \cdot W_{st} \cdot C_e + \beta_3 \cdot C_e + X'_{iest} \cdot \gamma + \delta_s + \phi_i + \nu_t + \varepsilon_{iest} \quad (8)$$

Dependent variable $P(Y_{iest} = 1)$ represents the probability that user i consumes (or separately purchases) episode e of series s at time t. W_{st} denotes the wait-time of series s at time t (log-transformed), and C_e is the normalized complementarity of episode e (based only on the text of episode $e-1$).[14] Control variables X_{iest} include episode-specific attributes such as length, age and position within the series. Fixed effects for series (δ_s), consumer (ϕ_i), and time (ν_t) account for unobserved heterogeneity and time trends.

The coefficients of interest are β_1, which captures the relationship between wait-time and the likelihood of consumption, and β_2, which measures how sequential complementarities moderate this relationship. The coefficient β_3 controls for baseline differences in consumption likelihood attributable to variation in complementarity across episodes. To maintain data consistency, we exclude observations where an episode is consumed without the immediately preceding episode.

Column (1) of Table 11 reports the results for the probability of consuming an episode

[14] Complementarity scores are normalized to have mean of 0 and standard deviation of 1 across all episodes.

(both paid and free). The coefficient on W_{st} is negative and significant, indicating that shorter wait-times are associated with a higher likelihood of consumption. Notably, the coefficient on the interaction term $W_{st} \cdot C_e$ is positive and significant. Together with the coefficient on C_e, it suggests that episodes with higher complementarities are more likely to be consumed under a given wait-time, with the positive effect of complementarities becoming more pronounced as wait-times increase.

Column (2) presents the results for the probability of purchasing an episode. The coefficient on W_{st} is positive and significant, indicating that shorter wait-times increase the substitutability between delayed (free) and immediate (paid) consumption, thereby cannibalizing revenues. The coefficient on $W_{st} \cdot C_e$ is negative and significant, suggesting that episodes with higher complementarities are less susceptible to revenue cannibalization when wait-times are reduced.

	(1) Pr(Consume)	(2) Pr(Purchase)
Intercept	5.508***	-1.614***
	(0.017)	(0.007)
W_{st}	-0.162***	0.404***
	(0.016)	(0.005)
$W_{st} \cdot C_e$	0.071***	-0.004***
	(0.005)	(0.001)
C_e	-0.057***	0.067***
	(0.005)	(0.001)
Eps Length	-0.002	-0.227***
	(0.010)	(0.002)
Eps Age	-0.025	0.031***
	(0.026)	(0.005)
Eps Position	0.267***	0.069***
	(0.011)	(0.002)
N Individual	121005	121005
N Obs	32286477	32286477

Note: Robust standard errors clustered at the consumer level in parentheses. All regressions include series, consumer and time fixed effects.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 11: Fixed effects logistic regression results

Figure 14 visualizes the estimated coefficients; the green dotted and yellow solid lines represent episodes with high and low complementarity values (greater than +/- 1 SD around

mean), respectively. While the logistic regression is descriptive, it highlights the dual forces of cannibalization and increased retention that collectively influence monetization within the consumer. Reducing wait-times encourages the consumer to consume the episodes, especially those with lower complementarities, and continue with the series. This progression increases the likelihood that the consumer will encounter episodes with high complementarities, where they are more inclined to pay for immediate access even under shorter wait-times.

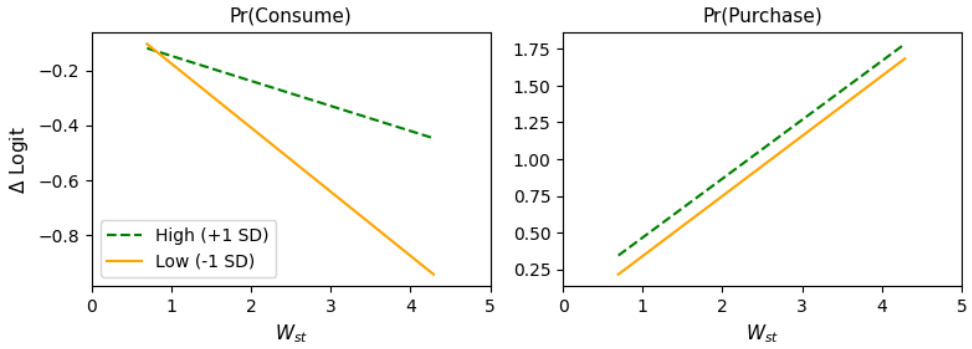


Figure 14: Logit$\Big[P(Y_{iest}=1)\Big]$ change by wait-time for high/low complementarity episodes

5.4 Discussion of the Potential Mechanisms

Our empirical findings suggest that sequential complementarities—where consuming one episode enhances the value of consuming the next—may play a pivotal role in driving increased aggregate and within-consumer consumption under shorter wait-times.

A defining property of sequential complementarities is that their value diminishes as the interval between consumption increases. As more time elapses since the consumption of the previous episode, the consumer's internal momentum wanes, and the memory of prior episodes fades, leading to a lower perceived value of the next episode (Murre and Dros, 2015; Anghelcev et al., 2021). Horvath et al. (2017) shows that viewers report lower enjoyment ratings for TV series as more time passes since their last viewing. Such an effect can also be observed in the addiction literature, which suggests that the utility of subsequent consumption decreases as the accumulated consumption capital dissipates over time (Becker

and Murphy, 1988; Heather and Vuchinich, 2003). This time-sensitive decline in valuation highlights the importance of timing strategies, including temporal versioning, in serialized media (Zhao et al., 2022; Godinho de Matos et al., 2023).

We find that the empirical results also align with this notion. The negative and significant coefficient on W_{st} in Column (1) of Table 11 suggests that longer wait-times reduce the likelihood of consumption, even when episodes are available for free. In contrast, shorter wait-times preserve sequential complementarities, keeping consumers engaged and increasing the likelihood of monetization through subsequent episodes.

The role of sequential complementarities is evident from Figure 12. Low-complementarity episodes are unlikely to be purchased regardless of wait-time, as they provide little incentive for immediate consumption. While reduced wait-times boost free consumption, they have minimal impact on paid consumption for these episodes. In contrast, high-complementarity episodes drive incremental monetization, as shorter waits sustain engagement, increasing the likelihood of purchasing subsequent episodes with strong complementarities.

Figure 13 indicates that wait-time reductions have a larger effect on the intensive margin for low-complementarity series than for high-complementarity series. For low-complementarity series, consumers are more likely to disengage under longer wait-times, as they encounter episodes with insufficient value—even when free. Shorter wait-times help them bypass these "obstacle episodes" before complementarities decay, creating additional purchase opportunities at subsequent episodes. This effect is less pronounced for high-complementarity series, as they naturally sustain engagement even under longer wait-times. This pattern is reflected in the positive and significant coefficient on $W_{st} \cdot C_e$ in Column (1) of Table 11.

Collectively, these results show that sequential complementarities are a key driver of increased monetization. While shorter wait-times may reduce the incentive to purchase, they also create positive spillovers across episodes, reducing churn, and inducing them to wait and/or purchase future episodes. Thus complementarities create additional monetization opportunities by retaining consumers longer over a broader product set.

6 Conclusion

Serialized content has emerged as a growing format on digital content platforms, capitalizing on its ability to foster consistent and recurring consumer engagement. Most platforms monetize this format using temporal versioning, a strategy that leverages time-based access to segment consumers. This paper examines how changing the wait-time between free and premium access affects user behavior and platform revenue for serialized media. Using data from a serialized book platform, we exploit an exogenous policy change where the platform reduced wait-times for a subset of series, it explores whether reducing the wait-time leads to net gains in paid unlocks or unintended cannibalization, and whether sequential complementarities—i.e., the strength of narrative connections between episodes measured using text data—amplify or mitigate these effects.

Our findings demonstrate that shorter wait-times increase overall consumption but yield non-monotonic effects on paid engagement. Moderate cuts often induce both new and existing readers to pay for immediate access, whereas extremely large cuts sometimes reduce users' willingness to pay if free access becomes too convenient. Textual analysis reveals that narrative continuity across episodes can substantially influence these outcomes: when cliffhangers or thematic hooks are strong, even a few hours of waiting are deemed intolerable by many readers, driving more paid unlocks. In contrast, users facing weaker inter-episode ties are content to wait out shorter delays, diminishing monetization gains.

By highlighting the interaction between time-based versioning and episode continuity, this research offers important implications for platforms and publishers. Rather than a one-size-fits-all wait policy, adjusting wait-times according to a series' narrative intensity can optimize revenue without sacrificing broad audience engagement. Moreover, systematically measuring textual complementarities through large language models or other natural language processing tools can help identify which titles or episodes will benefit most from shorter waits. Future work could investigate user-level or dynamic personalization of wait-times, as well as further explore the interplay between content attributes, consumption intensity, and

strategic release policies. Overall, the results underscore that temporal versioning in serialized contexts requires balancing user impatience with the narrative-driven urge to read on—a critical determinant of paid consumption in digital media.

We conclude with a discussion of limitations in our analysis. First, while we demonstrate the causal effect of changing wait-times, we cannot comment on the optimal wait-time, either for a series or specifically customized for a user. Second, our focus is on the short-term effects of wait-time reduction, as long-term effects are more challenging to identify. Future studies should investigate the long-term impacts of varying wait-times on platform-wide consumption and engagement. Third, whereas we examined the impact of wait time, it would be useful to incorporate the joint impact of other marketing variables like price.

Acknowledgments

The authors thank the participants at marketing seminars at Yale University.

Funding and Competing Interests

The authors have no funding or competing financial or non-financial interests to report in the subject matter presented in this manuscript.

References

Abadie, A. and G. W. Imbens (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics 29*(1), 1–11.

Abdellaoui, M. and E. Kemel (2014). Eliciting prospect theory when consequences are measured in time units:"time is not money". *Management Science 60*(7), 1844–1859.

Acemoglu, D. and A. Finkelstein (2008). Input and technology choices in regulated industries: Evidence from the health care sector. *Journal of Political Economy 116*(5), 837–880.

Anghelcev, G., S. Sar, J. D. Martin, and J. L. Moultrie (2021). Binge-watching serial video content: exploring the subjective phenomenology of the binge-watching experience. *Mass Communication and Society 24*(1), 130–154.

Angrist, J. D. and A. B. Krueger (1999). Empirical strategies in labor economics. In *Handbook of labor economics*, Volume 3, pp. 1277–1366. Elsevier.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics 10*(2), 150–161.

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

Becker, G. S. and K. M. Murphy (1988). A theory of rational addiction. *Journal of political Economy 96*(4), 675–700.

Berger, J., Y. D. Kim, and R. Meyer (2021). What makes content engaging? how emotional dynamics shape success. *Journal of Consumer Research 48*(2), 235–250.

Berger, J., W. W. Moe, and D. A. Schweidel (2023). What holds attention? linguistic drivers of engagement. *Journal of Marketing 87*(5), 793–809.

Borusyak, K., X. Jaravel, and J. Spiess (2024). Revisiting event-study designs: robust and efficient estimation. *Review of Economic Studies*, rdae007.

Bronnenberg, B. J., J.-P. Dubé, and R. E. Sanders (2020). Consumer misinformation and the brand premium: A private label blind taste test. *Marketing Science 39*(2), 382–406.

Butters, R. A., D. W. Sacks, and B. Seo (2022). How do national firms respond to local cost shocks? *American Economic Review 112*(5), 1737–1772.

Callaway, B., A. Goodman-Bacon, and P. H. Sant'Anna (2024). Difference-in-differences with a continuous treatment. Technical report, National Bureau of Economic Research.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics 134*(3), 1405–1454.

Choi, J., I. Chae, and F. Feinberg (2024). Wait for free: A consumption-decelerating promotion for serialized digital media. *Journal of Marketing Research*, 00222437241270194.

Cinelli, C. and C. Hazlett (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*(1), 39–67.

Clarke, D. (2017). Estimating difference-in-differences in the presence of spillovers. *Working Paper*.

Cook, L. D., M. E. Jones, T. D. Logan, and D. Rosé (2023). The evolution of access to public accommodations in the united states. *The Quarterly Journal of Economics 138*(1), 37–102.

Danaher, B., J. Hersh, M. D. Smith, and R. Telang (2020). The effect of piracy website blocking on consumer behavior. *Management Information Systems Quarterly 44*(2), 631–659.

Datta, H., G. Knox, and B. J. Bronnenberg (2018). Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery. *Marketing Science 37*(1), 5–21.

de Chaisemartin, C. and X. D'Haultfœuille (2020, September). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

Dehejia, R. H. and S. Wahba (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics 84*(1), 151–161.

Deng, Y., A. Lambrecht, and Y. Liu (2022). Spillover effects and freemium strategy in the mobile app market. *Management Science*.

Deshpande, M. and Y. Li (2019). Who is screened out? application costs and the targeting of disability programs. *American Economic Journal: Economic Policy 11*(4), 213–248.

Diamond, A. and J. S. Sekhon (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics 95*(3), 932–945.

Ely, J., A. Frankel, and E. Kamenica (2015). Suspense and surprise. *Journal of Political Economy 123*(1), 215–260.

Fong, H. and G. Gui (2024). Modeling story expectations to understand engagement: A generative framework using llms. *Columbia Business School Research Paper* (5053346).

Gardner, J. (2022). Two-stage differences in differences. *arXiv preprint arXiv:2207.05943*.

Ghose, A. and V. Todri-Adamopoulos (2016). Toward a digital attribution model. *MIS quarterly 40*(4), 889–910.

Godinho de Matos, M. and P. Ferreira (2020). The effect of binge-watching on the subscription of video on demand: Results from randomized experiments. *Information Systems Research 31*(4), 1337–1360.

Godinho de Matos, M., S. Mamadehussene, and P. Ferreira (2023). When less is more: Content strategies for subscription video on demand. *Available at SSRN 4352446*.

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

Guo, F. and Y. Liu (2023). The effectiveness of membership-based free shipping: An empirical investigation of consumers' purchase behaviors and revenue contribution. *Journal of Marketing 87*(6), 869–888.

Gynn, A. (2023). Serialized content: Why it's making a comeback for brands. https://www.thetilt.com/content/serialized-content-brands. Accessed on March 24, 2025.

Haigh, A., D. Apthorp, and L. A. Bizo (2021). The role of weber's law in human time perception. *Attention, Perception, & Psychophysics 83*, 435–447.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association 99*(467), 609–618.

Heather, N. and R. E. Vuchinich (2003). *Choice, behavioural economics and addiction*. Elsevier.

Heckman, J. J., H. Ichimura, J. A. Smith, and P. E. Todd (1998). Characterizing selection bias using experimental data. *Econometrica 66*(5), 1017–1098.

Horvath, J. C., A. J. Horton, J. M. Lodge, and J. A. Hattie (2017). The impact of binge watching on memory and perceived comprehension. *First Monday*.

Imai, K., I. S. Kim, and E. H. Wang (2023). Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science 67*(3), 587–605.

Jo, W., S. Sunder, J. Choi, and M. Trivedi (2020). Protecting consumers from themselves: Assessing consequences of usage restriction laws on online game usage and spending. *Marketing Science 39*(1), 117–133.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–292.

Kermode, F. (2000). *The sense of an ending: Studies in the theory of fiction with a new epilogue*. Oxford University Press.

Knight, S., M. D. Rocklage, and Y. Bart (2024). Narrative reversals and story success. *Science Advances 10*(34), eadl2013.

Krugmann, J. O. and J. Hartmann (2024). Sentiment analysis in the age of generative ai. *Customer Needs and Solutions 11*(1), 3.

Kumar, V. (2014). Making "freemium" work. *Harvard business review 92*(5), 27–29.

Lambrecht, A. and K. Misra (2017). Fee or free: When should firms charge for online content? *Management Science 63*(4), 1150–1165.

Leclerc, F., B. H. Schmitt, and L. Dube (1995). Waiting time and decision making: Is time like money? *Journal of consumer research 22*(1), 110–119.

Li, H., S. Jain, and P. Kannan (2019). Optimal design of free samples for digital products and services. *Journal of Marketing Research 56*(3), 419–438.

Linkis, S. T. (2021). *Serialization in Literature Across Media and Markets*. Routledge.

Lu, J., E. Bradlow, and J. Hutchinson (2019). Multiple dimensions of binging: The hidden costs and benefits. *Available at SSRN: https://ssrn.com/abstract=3493759*.

Lu, J., U. R. Karmarkar, and V. Venkatraman (2023). Planning-to-binge: Time allocation for future media consumption. *Journal of Experimental Psychology: Applied*.

Luan, J. Y. and K. Sudhir (2022). Optimal inter-release timing for sequentially released products. *Customer Needs and Solutions 9*(1-2), 25–46.

MarketingCharts (2019). Smartphones now account for 70% of us digital media time. https://www.marketingcharts.com/digital/mobile-phone-111093. Accessed April 10, 2025.

Michlin, M. (2011). More, more, more. contemporary american tv series and the attractions and challenges of serialization as ongoing narrative. *Mise au point. Cahiers de l'association française des enseignants et chercheurs en cinéma et audiovisuel* (3).

Mittell, J. (2006). Narrative complexity in contemporary american television. *The velvet light trap 58*(1), 29–40.

Murre, J. M. and J. Dros (2015). Replication and analysis of ebbinghaus' forgetting curve. *PloS one 10*(7), e0120644.

Narang, U. and V. Shankar (2019). Mobile app introduction and online and offline purchases and product returns. *Marketing Science 38*(5), 756–772.

Piper, A. and O. Toubia (2023). A quantitative study of non-linearity in storytelling. *Poetics 98*, 101793.

Poot, L. T. (2016). On cliffhangers. *Narrative 24*(1), 50–67.

Rathje, S., D.-M. Mirea, I. Sucholutsky, R. Marjieh, C. E. Robertson, and J. J. Van Bavel (2024). Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences 121*(34), e2308950121.

Reichl, P., S. Egger, R. Schatz, and A. D'Alconzo (2010). The logarithmic nature of qoe and the role of the weber-fechner law in qoe assessment. In *2010 IEEE International Conference on Communications*, pp. 1–5. IEEE.

Rosenbaum, P. R. (2002). *Overt bias in observational studies*. Springer.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Schlütz, D. M. (2016). Contemporary quality tv: The entertainment experience of complex serial narratives. *Annals of the International Communication Association 40:1*, 95–124.

Schweidel, D. A. and W. W. Moe (2016). Binge watching and advertising. *Journal of Marketing 80*(5), 1–19.

Shapiro, C. and H. R. Varian (1998). Versioning: the smart way to. *Harvard business review 107*(6), 107.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of econometrics 225*(2), 175–199.

Toubia, O., J. Berger, and J. Eliashberg (2021). How quantifying the shape of stories predicts their success. *Proceedings of the National Academy of Sciences 118*(26), e2011695118.

Varian, H. R. (2000). Versioning information goods. *Internet publishing and beyond: The economics of digital information and intellectual property*, 190–2002.

Wirz, D. S., A. Ort, B. Rasch, and A. Fahr (2023). The role of cliffhangers in serial entertainment: An experiment on cliffhangers' effects on enjoyment, arousal, and intention to continue watching. *Psychology of Popular Media 12*(2), 186.

Zeng, H. S., B. Danaher, and M. D. Smith (2022). Internet governance through site shutdowns: the impact of shutting down two major commercial sex advertising sites. *Management Science 68*(11), 8234–8248.

Zhang, S., T. Y. Chan, X. Luo, and X. Wang (2022). Time-inconsistent preferences and strategic self-control in digital content consumption. *Marketing Science 41*(3), 616–636.

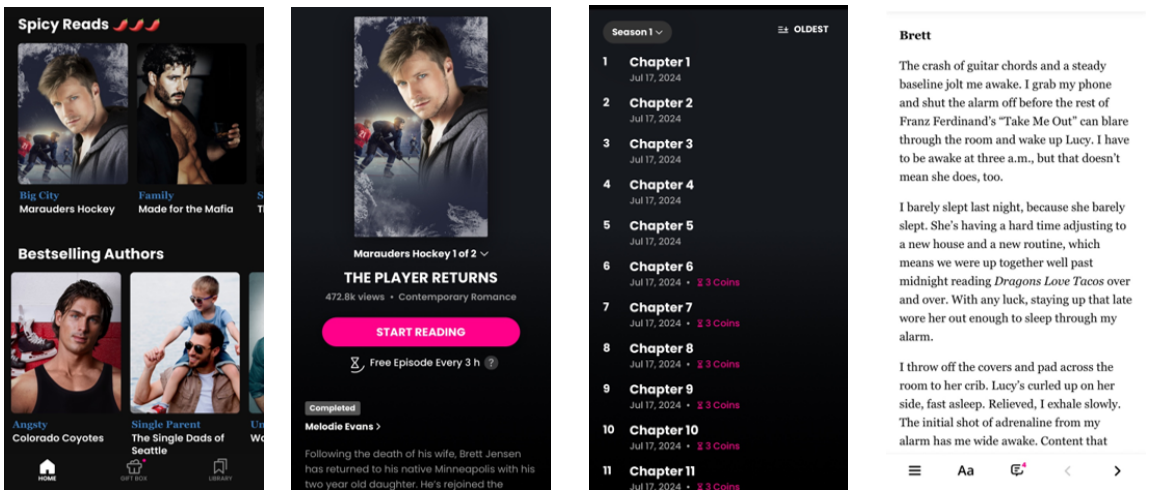
Zhao, C., N. Mehta, and M. Shi (2022). The consumption of serial media products and optimal release strategy. *Working Paper*.

Zillmann, D. (1995). Mechanisms of emotional involvement with drama. *Poetics 23*(1-2), 33–51.

A Web Appendix

A.1 App interface

Figure A.1 illustrates the user experience on the platform's mobile application. When a user opens the app, the home screen displays a list of available series (Figure A.1a). This screen is uniform across all users at any given time, and users can scroll through series or search for specific titles. Clicking on a series brings up additional details, such as the wait-time, genre, and a short description (Figure A.1b). For instance, the example shown features a contemporary romance series with a 3-hour wait-time. Below the description, episodes are listed in sequential order (Figure A.1c). In this example, the first five episodes are free, while subsequent episodes are offered under temporal versioning. An hourglass icon and the "3 Coins" label indicate that, starting with the sixth episode, users can either wait 3 hours to access the episode for free or pay 3 Coins to access it immediately. Episodes can only be consumed through the app and, once unlocked, remain accessible indefinitely at no additional cost. Each episode is presented in text form (Figure A.1d) and is approximately 1,500 words in length. The vast majority of readers complete an episode within 15 minutes.



(a) Home screen (b) Series overview (c) Episode list (d) Episode text

Figure A.1: User interface of the serialized novel platform

A.2 Using LLM to quantify sequential complementarities

The complementarity score of a given episode is quantified using the strength of the preceding episode's cliffhanger as a proxy. We leverage GPT-4o (version gpt-4o-2024-08-06) to analyze the full text of each episode and generate a single numerical score with one decimal precision, based on the provided prompt. To ensure consistency and reduce variability in the outputs, the model's temperature was set to zero.

System Instructions: You are a skilled literary analyst tasked with evaluating the strength of the cliffhanger in an episode of a serial novel.

Query: The strength of the cliffhanger is evaluated based on the following criteria:

- Tension and Suspense: How much does the scene make me anxious to find out what happens next?
- Emotional Investment: How much do I care about the characters affected by the cliffhanger?
- Surprise and Novelty: How unexpected or original is the cliffhanger?
- Stakes and Consequences: How significant are the potential outcomes for the characters or story?

Here is the episode text: {*EPISODE TEXT*}

For each criterion, provide a score from 0 = very low to 10 = very high, 0 being the weakest and 10 being the strongest. Use the entire range of the scale and be mindful of differences between episodes. Then, calculate the average of these scores to get the final overall score.

Example Response: 6.5

A.3 Prolific survey to validate complementarity measures

We validate our cliffhanger strength measures, obtained using a large language model, by testing whether humans perceive episodes with higher scores as having stronger cliffhangers than those with lower scores. We recruited 140 participants (72% female; $M_{age} = 39$) on Prolific Academic for a modest payment. Each participant was presented with two episodes—one high-scoring and one low-scoring—in a randomized order to control for potential order effects.

We created a curated pool of five low-score (below median score) episodes and five high-score (above median score) episodes. To construct this pool, we first selected the initial episode of each series, ensuring participants could follow the narrative without needing prior context. We then filtered for episodes below a specified length threshold and within the bottom ten percent of the scores. For each retained episode, we identified a high-score counterpart by selecting the most textually similar episode—measured via cosine similarity of text embeddings—among those above the median score and within the length requirement. This content-based matching minimizes confounding factors, ensuring differences in perceived cliffhanger strength were not driven by thematic or stylistic variation. After excluding episodes with excessively explicit content, five matched pairs were randomly selected. The final sample included five low-score episodes, with an average score of 3.5, and five high-score episodes, with an average score of 6.8. For reference, the full distribution of cliffhanger scores in our dataset ranges from 0.5 to 9, with a median of 6.5 (Figure 3).

After each episode, participants answered reading comprehension questions to confirm they had carefully engaged with the content; two participants who failed these questions were excluded from the analysis. In addition, copy-and-paste functionality was disabled to prevent the use of language models for generating responses or evaluating cliffhangers.

Upon reading both episodes, participants were first asked which episode they enjoyed more, or whether they liked/disliked both episodes equally. This question was designed to control for the potential confound between overall enjoyment and objective evaluation of cliffhanger strength. Next, participants were asked to select the episode with the stronger

cliffhanger (binary choice) and were explicitly instructed to base their choice solely on the cliffhanger's strength, independent of their personal preference. To ensure consistency, we provided the same evaluation criteria previously given to the LLM in Section A.2. Finally, participants were asked to provide an open-ended explanation for their choice.

Of the 138 participants, 93 (67%) identified the high-score episode as having the stronger cliffhanger. A proportion z-test ($H_0 : p = 0.5$) confirmed that this result was significantly different from chance ($p < 0.001$). On the initial preference question, 80 participants (58%) reported enjoying the high-score episode more, while 37 (27%) preferred the low-score episode, suggesting a positive correlation between cliffhanger strength and enjoyment. These findings provide strong support for the validity of the LLM-generated cliffhanger scores, indicating that they closely align with human perception.

A.4 Evidence on randomness of post-reduction wait-time assignment

We empirically assess the platform's claim that, conditional on selection for a wait-time reduction, post-reduction wait-times were assigned randomly. Table A.1 presents the results of regressing post-reduction wait-time on the covariates used for propensity score matching. With the exception of pre-treatment wait-time—where a correlation is expected since post-reduction wait-time is derived from it—none of the coefficients are statistically significant ($p > 0.1$). These findings support the firm's claim that post-reduction wait-times were assigned randomly, conditional on treatment.

	Post-reduction Wait-time	
	Coefficient	Standard Error
Intercept	1.578	(3.661)
T1 Purchased Eps	-0.657	(0.424)
T2 Purchased Eps	-0.673	(0.468)
T3 Purchased Eps	-0.503	(0.496)
T4 Purchased Eps	0.262	(0.411)
T1 Waited Eps	0.990	(1.743)
T2 Waited Eps	-2.033	(2.398)
T3 Waited Eps	1.715	(2.729)
T4 Waited Eps	1.132	(2.512)
T1 Promotion	1.516	(1.470)
T2 Promotion	-1.809	(1.585)
T3 Promotion	-0.158	(0.439)
T4 Promotion	0.563	(0.433)
T1 Consumers	-0.673	(2.051)
T2 Consumers	3.340	(2.829)
T3 Consumers	-2.064	(3.050)
T4 Consumers	-1.017	(2.758)
No. Free Eps	0.005	(0.507)
No. WFF Eps	-0.143	(0.391)
No. Paid Eps	0.211	(0.363)
Series Age	0.010	(0.281)
Wait-time	0.076***	(0.022)
N Obs	211	
Adj. R^2	0.367	

Note: Includes genre and time fixed effects.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A.1: Post-reduction wait-time regressed on covariates used for propensity score matching

A.5 Full estimation results

We present the full estimation results of the DiD analysis from Section 4.5, including the control variables. Due to space constraints, we show the results of the parametric specification with better model fit based on AIC (Equation 4) and the semi-parametric specification (Equation 6).

	Series-level Aggregate Consumption						Extensive Margins		Intensive Margins					
	Total Consumption		Paid Consumption		Free Consumption		Total Consumers		Total Consumption		Paid Consumption		Free Consumption	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
$T \cdot \mathbb{1}(R > 0)$	0.201		0.484***		0.174		0.311***		0.310		0.336***		0.064	
	(0.229)		(0.138)		(0.237)		(0.113)		(0.194)		(0.109)		(0.240)	
$T \cdot R$	0.295		-0.130**		0.342*		-0.118***		-0.292*		-0.452***		-0.007	
	(0.195)		(0.052)		(0.203)		(0.040)		(0.156)		(0.090)		(0.188)	
$T \cdot R^2$	-0.041				-0.031				0.110***		0.113***		0.056	
	(0.040)				(0.042)				(0.032)		(0.019)		(0.038)	
$T \cdot \mathbb{1}(R \in \mathcal{L}_1)$		0.424***		0.359***		0.440***		0.159		0.202**		0.162***		0.090
		(0.114)		(0.124)		(0.120)		(0.102)		(0.082)		(0.059)		(0.130)
$T \cdot \mathbb{1}(R \in \mathcal{L}_2)$		0.562***		0.210**		0.688***		0.105		0.063		-0.157***		0.148
		(0.094)		(0.099)		(0.098)		(0.084)		(0.097)		(0.046)		(0.117)
$T \cdot \mathbb{1}(R \in \mathcal{L}_3)$		0.734***		0.147**		0.918***		-0.020		0.448***		-0.006		0.566***
		(0.054)		(0.067)		(0.056)		(0.045)		(0.043)		(0.024)		(0.049)
$T \cdot \mathbb{1}(R \in \mathcal{L}_4)$		0.683***		-0.179		0.994***		-0.155**		0.762***		0.304***		0.786***
		(0.111)		(0.119)		(0.130)		(0.077)		(0.073)		(0.050)		(0.086)
N Free Eps	0.167	0.134	0.172	0.177	0.148	0.104	0.155	0.180	0.214***	0.222***	-0.325***	-0.303***	0.420***	0.416***
	(0.157)	(0.159)	(0.155)	(0.158)	(0.170)	(0.175)	(0.123)	(0.125)	(0.062)	(0.075)	(0.049)	(0.055)	(0.068)	(0.075)
N WFF Eps	1.537***	1.501***	1.323***	1.317***	1.436***	1.404***	1.391***	1.409***	-0.413*	-0.460*	-1.526***	-1.559***	0.527**	0.484*
	(0.323)	(0.329)	(0.346)	(0.351)	(0.300)	(0.307)	(0.262)	(0.267)	(0.236)	(0.254)	(0.175)	(0.181)	(0.264)	(0.279)
N Paid Eps	0.141	0.155	0.280**	0.274**	0.084	0.097	-0.048	-0.056	-0.005	-0.014	0.198***	0.164***	-0.207	-0.190
	(0.129)	(0.128)	(0.128)	(0.128)	(0.144)	(0.143)	(0.101)	(0.100)	(0.079)	(0.091)	(0.073)	(0.057)	(0.156)	(0.159)
Promo	0.087***	0.087***	0.112***	0.112***	0.074***	0.074***	0.128***	0.128***	0.102***	0.105***	-0.019	-0.017	0.181***	0.183***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)	(0.013)	(0.013)	(0.033)	(0.033)	(0.018)	(0.018)	(0.043)	(0.043)
T7 Promo	0.129***	0.129***	0.113***	0.114***	0.136***	0.136***	0.096***	0.096***						
	(0.016)	(0.016)	(0.014)	(0.014)	(0.017)	(0.017)	(0.014)	(0.014)						
N Obs	64064	64064	64064	64064	64064	64064	64064	64064	740722	740722	740722	740722	740722	740722
Adj. R^2	0.072	0.073	0.020	0.020	0.120	0.120	0.076	0.075	0.006	0.006	0.001	0.001	0.008	0.008
AIC	99882	99872	155526	155526	76689	76694	59594	59627	1800726	1800751	1981495	1981423	1919075	1919038

Note: Robust standard errors clustered at the series level in parentheses. All regressions include group-series and group-time fixed effects.

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A.2: Full estimation results of the main DiD analysis

A.6 Treatment effect on episode-level aggregate consumption

To ensure the robustness of our results, we replicate the analysis in Section 4.5 with consumption aggregated at the episode level. While the series-level analysis provides a broader perspective on treatment effects, aggregating across all episodes within a series may mask certain within-series dynamics, even after controlling for the number of episodes by access type. Episode-level analysis allows us to additionally control for variations in characteristics specific to individual episodes, including their position within the series, access type, release date and word count. The results, presented in Table A.3, are qualitatively consistent with the series-level findings. Columns (1)-(3) show that total consumption increases with treatment intensity, while Columns (4)-(6) reveal that wait-time reductions increase paid consumption up to a certain threshold magnitude. Columns (7)-(9) show the positive treatment effect on free consumption, increasing with treatment intensity. The treatment effects are visualized in Figure A.2, demonstrating consistency across the two model specifications. However, the level-specific estimate for $\mathcal{L}_4$ on paid consumption is notably more negative. Overall, the results suggest that the platform can garner higher series revenues by allowing quicker free consumption, with the important caveat that excessively short wait-times may lead to adverse effects.

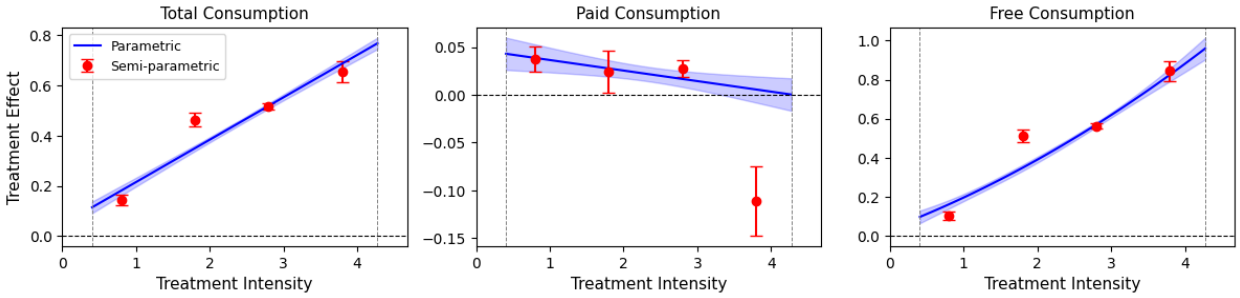


Figure A.2: Treatment effect on episode-level aggregate consumption by treatment intensity. Blue line and red dots represent estimates from the parametric (Eq. 4) and semi-parametric (Eq. 6) specifications, respectively. Shaded region and red error bars indicate 95% confidence intervals.

	Total Consumption			Paid Consumption			Free Consumption		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$T \cdot \mathbb{1}(R > 0)$	0.046***	0.060***		0.047***	0.041**		-0.032**	0.037	
	(0.014)	(0.023)		(0.010)	(0.016)		(0.015)	(0.025)	
$T \cdot R$	0.169***	0.154***		-0.011***	-0.004		0.218***	0.143***	
	(0.005)	(0.023)		(0.004)	(0.018)		(0.006)	(0.026)	
$T \cdot R^2$		0.003			-0.002			0.017***	
		(0.005)			(0.004)			(0.006)	
$T \cdot \mathbb{1}(R \in \mathcal{L}_1)$			0.144***			0.037***			0.104***
			(0.010)			(0.007)			(0.010)
$T \cdot \mathbb{1}(R \in \mathcal{L}_2)$			0.464***			0.024**			0.513***
			(0.014)			(0.011)			(0.016)
$T \cdot \mathbb{1}(R \in \mathcal{L}_3)$			0.516***			0.027***			0.563***
			(0.006)			(0.004)			(0.007)
$T \cdot \mathbb{1}(R \in \mathcal{L}_4)$			0.655***			-0.111***			0.843***
			(0.021)			(0.019)			(0.026)
N Obs	4152204	4152204	4152204	4152204	4152204	4152204	4152204	4152204	4152204
Adj. R^2	0.051	0.051	0.051	0.048	0.048	0.048	0.152	0.152	0.152
AIC	9974230	9974231	9974485	8910527	8910528	8910394	9980952	9980933	9981305

Note: Robust standard errors clustered at the episode level in parentheses. All regressions include observed control variables, group-series and group-time fixed effects.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A.3: Treatment effect on episode-level daily consumption

A.7 Additional robustness checks

We demonstrate the robustness of our results using several checks: alternative windows of two and eight weeks around the treatment (Section A.7.1), pseudo-treatment effects (Section A.7.2), subsample analyses using (1) series with no promotional activities and (2) treated series with 1-hour post-treatment wait-times (Section A.7.3), and sensitivity analysis to account for potential unobserved confounders (Section A.7.4). For conciseness and interpretability, we report the level-specific treatment effect estimates based on Equation 6. We also confirm that our main results align with treatment effects estimated using a fully non-parametric method proposed by Callaway et al. (2024) (Section A.7.5).

A.7.1 Alternative time window around treatment

The main analysis focuses on a four-week window around the treatment, under the assumption that any changes to consumption patterns within this period can be attributed to the change in wait-times. To assess the robustness of our results to alternative time frames, we re-estimate the treatment effects using two- and eight-week windows around the treatment.

The results, reported in Table A.4 with series-level aggregate consumption as the dependent variable, remain qualitatively consistent with the main findings.

	(1) Two-week Window			(2) Eight-week Window		
DV: Consumption	Total	Paid	Free	Total	Paid	Free
$T \cdot \mathbb{1}(R \in \mathcal{L}_1)$	0.429***	0.338**	0.450***	0.256**	0.252**	0.265**
	(0.110)	(0.147)	(0.109)	(0.125)	(0.126)	(0.126)
$T \cdot \mathbb{1}(R \in \mathcal{L}_2)$	0.583***	0.151*	0.770***	0.509***	0.198*	0.627***
	(0.067)	(0.086)	(0.073)	(0.110)	(0.104)	(0.111)
$T \cdot \mathbb{1}(R \in \mathcal{L}_3)$	0.847***	0.197***	1.045***	0.632***	0.122*	0.797***
	(0.045)	(0.070)	(0.045)	(0.064)	(0.070)	(0.066)
$T \cdot \mathbb{1}(R \in \mathcal{L}_4)$	1.047***	0.064	1.371***	0.292**	-0.475***	0.637***
	(0.102)	(0.142)	(0.112)	(0.142)	(0.133)	(0.163)
N Obs	32032	32032	32032	128128	128128	128128
Adj. R^2	0.083	0.012	0.167	0.075	0.028	0.108

Note: Robust standard errors clustered at the series level in parentheses. All regressions include observed control variables, group-series and group-time fixed effects.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A.4: Treatment effect on series-level daily consumption using alternative time windows

A.7.2 Pseudo-treatment effects

We assess the possibility that the estimated treatment effects may be driven by spurious correlations through falsification tests using pseudo-treatment indicators and pseudo-treatment dates. For pseudo-treatment indicators, we randomly assign control series as treated within each series group and re-estimate the model, excluding the actual treated series. Since these pseudo indicators do not reflect actual wait-time reductions, the estimated treatment effects should not be significant (Ghose and Todri-Adamopoulos, 2016; Jo et al., 2020). For pseudo-treatment dates, we shift the treatment date to three weeks prior to the actual reduction date. As this adjusted time frame excludes the true treatment period, the estimates should again be insignificant. Table A.5 confirms that the estimated treatment effects are not statistically significant under either falsification test, supporting the validity of our findings and ruling out concerns that they may be statistical artifacts of the model specification.

	(1) Pseudo Treatment Indicator			(2) Pseudo Treatment Date		
DV: Consumption	Total	Paid	Free	Total	Paid	Free
$T \cdot \mathbb{1}(R \in \mathcal{L}_1)$	-0.074	-0.130	-0.059	-0.069	-0.198	-0.026
	(0.132)	(0.123)	(0.134)	(0.140)	(0.147)	(0.136)
$T \cdot \mathbb{1}(R \in \mathcal{L}_2)$	0.143	0.112	0.165*	0.108	0.119	0.089
	(0.103)	(0.111)	(0.097)	(0.082)	(0.108)	(0.077)
$T \cdot \mathbb{1}(R \in \mathcal{L}_3)$	0.017	0.013	0.028	0.014	0.038	0.007
	(0.050)	(0.063)	(0.046)	(0.050)	(0.067)	(0.045)
$T \cdot \mathbb{1}(R \in \mathcal{L}_4)$	0.097	0.065	0.189	-0.083	0.016	-0.080
	(0.162)	(0.192)	(0.152)	(0.093)	(0.134)	(0.092)
N Obs	116312	116312	116312	63989	63989	63989
Adj. R^2	0.058	0.026	0.076	0.063	0.028	0.088

Note: Robust standard errors clustered at the series level in parentheses. All regressions include observed control variables, group-series and group-time fixed effects.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A.5: Falsification tests using pseudo treatment indicators and dates

A.7.3 Subsample analysis

Next, we examine whether the estimated treatment effect could be influenced by the platform's strategic timing of wait-time reductions and promotional activities. If promotions were run in anticipation of, or concurrently with, the wait-time reduction, the estimated treatment effect might be confounded by the impact of promotions. Although we explicitly control for promotions in all our analyses, we further address this concern by re-estimating the model after excluding any series (treated or non-treated) that had active promotions during the four-week window.

We also assess robustness using a subsample of treated series with identical post-reduction wait-times. As shown in Table 2, treated series exhibit variation in both pre- and post-reduction wait-times. While our main analysis estimates the effect as a function of treatment intensity—defined as the proportional reduction in wait-times—to account for these differences, we further alleviate potential concerns by restricting the analysis to a subsample of 159 treated series where wait-times were reduced to one hour. The results, presented in Table A.6, confirm that the treatment effects remain robust across both subsample analyses.

	(1) No Promotions			(2) 1-hour Post-treatment Wait-time		
DV: Consumption	Total	Paid	Free	Total	Paid	Free
$T \cdot \mathbb{1}(R \in \mathcal{L}_1)$	0.424***	0.387***	0.428***	0.359*	0.329	0.364*
	(0.114)	(0.131)	(0.118)	(0.189)	(0.202)	(0.196)
$T \cdot \mathbb{1}(R \in \mathcal{L}_2)$	0.491***	0.160*	0.622***	0.696***	0.347**	0.833***
	(0.084)	(0.095)	(0.090)	(0.187)	(0.168)	(0.197)
$T \cdot \mathbb{1}(R \in \mathcal{L}_3)$	0.704***	0.136**	0.870***	0.710***	0.090	0.914***
	(0.055)	(0.069)	(0.055)	(0.056)	(0.067)	(0.058)
$T \cdot \mathbb{1}(R \in \mathcal{L}_4)$	0.716***	-0.161	1.046***	0.683***	-0.179	0.994***
	(0.121)	(0.125)	(0.141)	(0.111)	(0.119)	(0.130)
N Obs	56448	56448	56448	48104	48104	48104
Adj. R^2	0.039	0.004	0.080	0.073	0.018	0.125

Note: Robust standard errors clustered at the series level in parentheses. All regressions include observed control variables, group-series and group-time fixed effects.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A.6: Subsample analyses using series with no promotions and 1-hour post-treatment wait-time

A.7.4 Sensitivity analysis

To assess the robustness of our treatment effect estimates to unobservable confounding, we conduct a sensitivity analysis following the method proposed by Cinelli and Hazlett (2020). While propensity score matching balances observable characteristics, unobserved factors may still bias our estimates. Cinelli and Hazlett (2020) extends the omitted variable bias framework to provide a scale-free, partial R^2-based approach that does not require assumptions on the functional form of treatment assignment or the distribution of unobserved confounders. This method quantifies the strength of an unobserved confounder—relative to an observed benchmark covariate—that would be required to nullify the estimated treatment effect.

We use the number of WFF episodes in the series, as the benchmark covariate because it has the highest explanatory power for treatment assignment among the control variables. Table A.7 reports the treatment effect estimates across a range of k-values, where k represents the relative strength of the unobserved confounder compared to the benchmark. Even in extreme cases where an unobserved confounder is assumed to be up to 2.5 times as strong as the benchmark in explaining treatment assignment, the estimated treatment effects remain

statistically significant and qualitatively unchanged compared to the baseline estimates ($k = 0$). These results indicate that our findings are robust to potential biases from unobserved confounding, as such confounders would need to be implausibly strong to undermine our conclusions.

	Relative Explanatory Power of Unobserved Confounders					
	$k = 0$	$k = 0.5$	$k = 1.0$	$k = 1.5$	$k = 2.0$	$k = 2.5$
DV: Total Consumption						
$T \cdot \mathbb{1}(R \in \mathcal{L}_1)$	0.424***	0.421***	0.418***	0.414***	0.411***	0.408***
	(0.114)	(0.114)	(0.114)	(0.114)	(0.114)	(0.114)
$T \cdot \mathbb{1}(R \in \mathcal{L}_2)$	0.562***	0.528***	0.493***	0.458***	0.423***	0.387***
	(0.094)	(0.094)	(0.095)	(0.095)	(0.096)	(0.097)
$T \cdot \mathbb{1}(R \in \mathcal{L}_3)$	0.734***	0.729***	0.724***	0.720***	0.715***	0.710***
	(0.054)	(0.054)	(0.054)	(0.054)	(0.054)	(0.054)
$T \cdot \mathbb{1}(R \in \mathcal{L}_4)$	0.683***	0.680***	0.677***	0.674***	0.671***	0.668***
	(0.111)	(0.111)	(0.111)	(0.111)	(0.111)	(0.111)
DV: Paid Consumption						
$T \cdot \mathbb{1}(R \in \mathcal{L}_1)$	0.359***	0.356***	0.353***	0.350***	0.348***	0.345***
	(0.124)	(0.124)	(0.124)	(0.124)	(0.124)	(0.124)
$T \cdot \mathbb{1}(R \in \mathcal{L}_2)$	0.210**	0.180*	0.150	0.120	0.089	0.058
	(0.099)	(0.100)	(0.101)	(0.101)	(0.102)	(0.103)
$T \cdot \mathbb{1}(R \in \mathcal{L}_3)$	0.147**	0.143**	0.138**	0.133**	0.128*	0.123*
	(0.067)	(0.067)	(0.067)	(0.067)	(0.067)	(0.067)
$T \cdot \mathbb{1}(R \in \mathcal{L}_4)$	-0.179	-0.176	-0.173	-0.171	-0.168	-0.166
	(0.119)	(0.119)	(0.119)	(0.119)	(0.119)	(0.119)
DV: Free Consumption						
$T \cdot \mathbb{1}(R \in \mathcal{L}_1)$	0.440***	0.437***	0.433***	0.430***	0.427***	0.423***
	(0.120)	(0.119)	(0.119)	(0.119)	(0.119)	(0.119)
$T \cdot \mathbb{1}(R \in \mathcal{L}_2)$	0.688***	0.653***	0.616***	0.580***	0.543***	0.505***
	(0.098)	(0.099)	(0.099)	(0.100)	(0.100)	(0.101)
$T \cdot \mathbb{1}(R \in \mathcal{L}_3)$	0.918***	0.913***	0.908***	0.903***	0.898***	0.893***
	(0.056)	(0.056)	(0.056)	(0.056)	(0.056)	(0.056)
$T \cdot \mathbb{1}(R \in \mathcal{L}_4)$	0.994***	0.991***	0.987***	0.984***	0.980***	0.977***
	(0.130)	(0.130)	(0.130)	(0.130)	(0.130)	(0.130)

Note: Robust standard errors clustered at the series level in parentheses. All regressions include observed control variables, group-series and group-time fixed effects.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A.7: Robustness to unobserved confounders: Cinelli-Hazlett bounds

A.7.5 Non-parametric estimation

To further validate the robustness of our results, we employ a non-parametric estimation approach based on the method proposed by Callaway et al. (2024). In the presence of con-

tinuous treatments, the authors of the paper argue that parametric DiD model specifications such as Equation 4 have two key limitations in estimating treatment effects. First, the estimated effect may be "contaminated" by selection bias, as the average outcomes for units receiving different treatment intensities might differ even if both were assigned the same intensity. The standard parallel trends assumption is insufficient to rule this out as it only pertains to the potential outcomes of untreated units. Second, the estimates are weighted averages of individual treatment effects with non-intuitive weights. These weights disproportionately favor treatment effects near the middle of the intensity range, diverging from the actual distribution of intensities. It is important to note that while these limitations apply to the parametric specification, our semi-parametric specification (Equation 6) avoids such biases by estimating treatment effects separately for discrete intensity levels.

To address these concerns, the authors propose a non-parametric approach that directly estimates treatment effects without imposing any functional form, offering a flexible and robust alternative. This method avoids the biases discussed above and provides clear causal interpretations of treatment effects across the full range of intensities. The estimation method relies on the strong parallel trends assumption, which is stricter than the standard parallel trends assumption. Formally, the assumption requires that for all treatment intensities $r \in \mathcal{R}$, $\mathbb{E}\Big[Y_{t=2}(r) - Y_{t=1}(0)\Big] = \mathbb{E}\Big[Y_{t=2}(r) - Y_{t=1}(0) \mid R = r\Big]$. This condition implies that the evolution of outcomes for one treatment intensity group must reflect what would have happened to other groups had they instead received the same intensity. In essence, it extends the standard parallel trends assumption from comparisons between treated and untreated groups to comparisons among treated groups with varying treatment intensities. When the strong parallel trends assumption is satisfied, the selection bias is eliminated.

In our analysis, we address potential violations of this assumption through propensity matching to ensure that treatment selection is random conditional on observed covariates. Conditional on receiving treatment, the platform indicated that post-reduction wait-times were randomly assigned, a claim we empirically validate in Section A.4. Additionally, we

demonstrate in Table 8 that there are no significant pre-treatment deviations in outcome trends moderated by treatment intensity, providing further evidence in support of the strong parallel trends assumption.

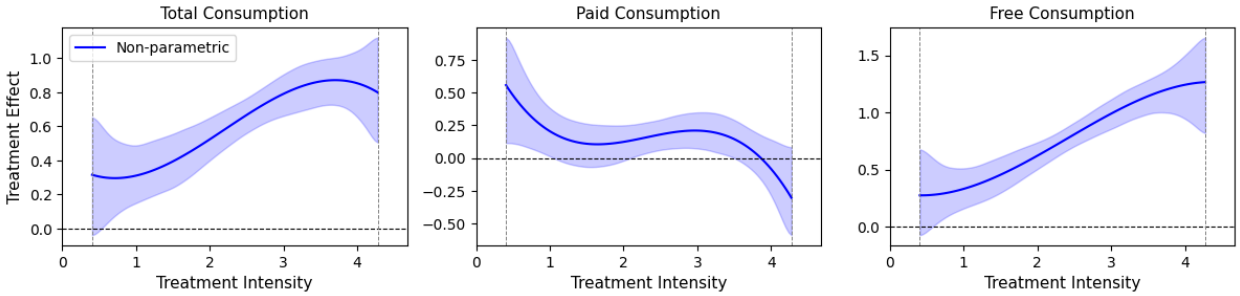
The non-parametric estimation procedure proceeds in the following steps:

1. Pick a family of basis functions $\psi^K(r)$
2. Pick the sieve dimension $\hat{K} \in \mathcal{K} = \{(2^k+3) : k \in \mathbb{N} \cup 0\}$, i.e., how many transformations of R to include in the regression. Including too many terms risks over-fitting and imprecise estimates, while including too few terms risks failing to capture heterogeneity.
3. Given sieve dimension $\hat{K}$, estimate the OLS coefficients, $\hat{\beta}_K$, of the "transformed outcome" $\Delta Y - \mathbb{E}_n\Big[\Delta Y \mid R = 0\Big]$, onto the K-dimensional B-spline $\psi_K(R)$ in the subsample of units with non-zero treatment intensity.
4. The proposed nonparametric estimator is given by $\widehat{ATE}_{\hat{K}}(r) = \Big(\psi^{\hat{K}}(r)\Big)\hat{\beta}_{\hat{K}}$

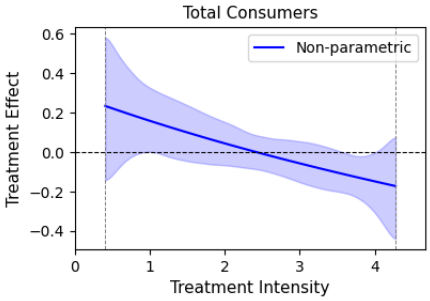
Following the suggestions from the paper, we use dyadic cubic B-splines as basis functions and set $\hat{K} = 4$, the lowest feasible value, to maintain parsimony. We refer readers to the original paper for additional details on the estimation method.

Figure A.3 presents the treatment effects estimated using the non-parametric method. The findings align closely with the main DiD results, offering additional validation. At the series-level, the treatment effect on paid consumption (Figure A.3a, middle panel) is positive for low treatment intensities but declines as intensity increases, while the treatment effect on free consumption (Figure A.3a, right panel) is consistently positive and grows monotonically with treatment intensity. For the extensive margin (Figure A.3b), the treatment effect starts positive at low intensities but declines linearly. On the intensive margin, we observe a U-shaped relationship for paid consumption per consumer (Figure A.3c, middle panel): the treatment effect is positive at low intensities, turns negative at moderate intensities, and becomes positive again at higher intensities. For free consumption per consumer (Figure

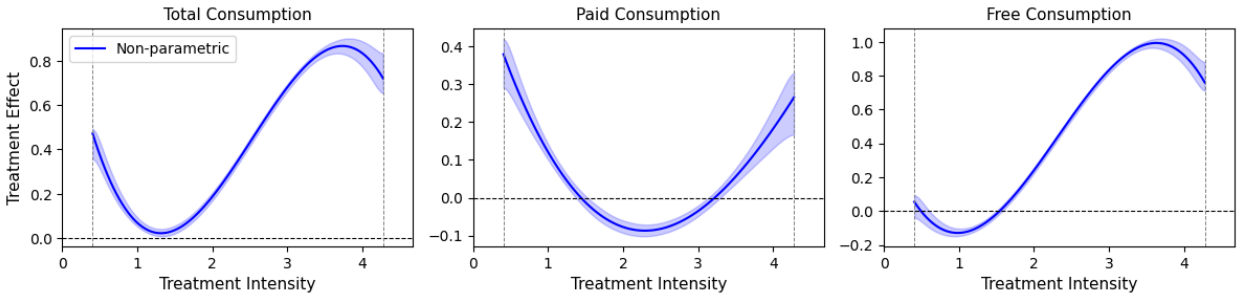
A.3c, right panel), the treatment effect is negligible or slightly negative at low intensities but increases steadily as the treatment intensity grows.



(a) Series-level aggregate consumption



(b) Extensive Margins: Breadth of Series Consumption



(c) Intensive Margins: Depth of Series Consumption

Figure A.3: Non-parametric treatment effect estimates using the method proposed by Callaway et al. (2024). Shaded regions indicate 95% confidence interval derived from bootstrap resampling.

A.8 Additional estimation details for mechanism check

	(1) Total Consumption	(2) Paid Consumption	(3) Free Consumption
Below Median			
$T \cdot \mathbb{1}(R > 0)$	0.015	-0.012	-0.073
	(0.060)	(0.027)	(0.066)
$T \cdot R$	0.048	0.004	0.072
	(0.054)	(0.010)	(0.061)
$T \cdot R^2$	0.043***		0.054***
	(0.011)		(0.013)
Above Median			
$T \cdot \mathbb{1}(R > 0)$	0.022	0.082***	-0.077*
	(0.036)	(0.014)	(0.040)
$T \cdot R$	0.137***	-0.024***	0.180***
	(0.036)	(0.005)	(0.040)
$T \cdot R^2$	0.008		0.013
	(0.008)		(0.009)
N Obs	2895844	2895844	2895844
Adj. R^2	0.046	0.046	0.163

Note: Robust standard errors clustered at the episode level in parentheses. All regressions include observed control variables, group-series and group-time fixed effects. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A.8: Treatment effect on episode-level aggregate consumption based on sequential complementarities

	(1) Total Consumption	(2) Paid Consumption	(3) Free Consumption
Below Median			
$T \cdot \mathbb{1}(R > 0)$	0.814**	0.641***	0.663
	(0.407)	(0.132)	(0.514)
$T \cdot R$	-0.562*	-0.657***	-0.297
	(0.312)	(0.118)	(0.377)
$T \cdot R^2$	0.148**	0.150***	0.089
	(0.058)	(0.024)	(0.069)
Above Median			
$T \cdot \mathbb{1}(R > 0)$	0.314*	0.277**	0.109
	(0.164)	(0.111)	(0.222)
$T \cdot R$	-0.417***	-0.430***	-0.227
	(0.148)	(0.111)	(0.190)
$T \cdot R^2$	0.142***	0.106***	0.113***
	(0.033)	(0.027)	(0.040)
N Obs.	740722	740722	740722
Adj. R^2	0.006	0.001	0.008

Note: Robust standard errors clustered at the series level in parentheses. All regressions include observed control variables, group-series and group-time fixed effects.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table A.9: Treatment effect on consumption per consumer based on sequential complementarities