Serialized Information Goods:

Access Delay as a Monetization Lever

Peter S. Lee, Vineet Kumar, K. Sudhir

February 8, 2024

Abstract

Serialization whereby a unified content is partitioned into smaller installments has become a prevalent form of publication for digital platforms offering information goods. Recently, these platforms have begun to adopt a novel monetization strategy called "wait-for-free" (WFF) that leverages the trade-off between time and money by allowing consumers the option of free, but delayed access. This paper investigates the impact of shortening wait-times on the profitability of serialized information goods, a counter-intuitive approach that would encourage free consumption over purchases. We highlight three conceptual features unique to serialized information goods: defined sequence of consumption, complementarity between episodes and diminishing value of complementarity over time. Based on these features, we hypothesize that reducing wait-times may actually lead to higher revenues by increasing the expected valuation of episodes under uncertainty of consumer preferences. We leverage a natural experiment from a serial fiction platform to identify whether the wait-time reduction stimulates or hurts demand for paid consumption. We estimate the effect under a difference-in-difference framework, while addressing potential selection issues using a matching-based approach. We find that reduction in wait-time increases both free and paid consumption by 115% and 23%, respectively. We then conduct a battery of robustness checks rule out spurious correlations. Finally, we propose and conduct three empirical tests to validate our hypothesis and find that more consumers progress further in the series at a faster rate, which together explains the increase in revenues.

1 Introduction

> "In a highly competitive environment – fighting to occupy users' free time – producers and distributors of content depend more than ever on the engagement of the services they offer. The sustained, consistent and recurring consumption of narrative series has demonstrated its effectiveness above any other format in building up that engagement."
>
> (*Storytel 2018 report*, Link)

Serialized information goods refer to products such as books, TV shows or educational courses that consist of multiple episodes with a continuous plot under a single title. Modern advancements in digital media, such as the emergence of ebook and streaming platforms, have paved way for serialization as the dominant mode of publication across product types. Partitioning a unified content in short episodes is also well received as an effective format that fits the behavioral consumer trends such as growing media consumption on mobile devices and the diminishing attention span of the users (Marketing Charts, 2019; Speaking of Psychology, 2023). Such popularity amongst users and publishers alike has led to a remarkable growth of the industry. The largest platform for serialized fiction novels, Wattpad, is reported to have over 80 million readers, as well as the leading serialized comics platform, Webtoon, boasting over 85 million users. The video streaming giants Netflix, Amazon Prime Video and Disney+ together serve more than 600 million subscribers.

Researchers in literary studies and semiology highlight the unique structural features of serialized information goods that affect people's consumption decisions. First, consumers engage with each episode according to a predetermined sequence, essential for comprehending the overarching plot. Mittell (2006) calls out the narrative complexity of modern TV series, where devices such as analepses or alterations in chronology are implemented to allow the audience to slowly build their understanding of and become hooked to the storyline that gradually unfolds. This sequential engagement is not just a narrative device but also influences economic decisions, as the consumption of each episode informs the decision to proceed to the next. Second, episodes exhibit directed complementarities: consuming one episode enhances the value of subsequent ones. Each episode in a series strikes a balance between repetition and variation, diluting the idea of the ending – while adjacent episodes share structural and content similarities, they also introduce new elements that expand on the earlier episode (Eco, 1990; Kermode, 2000). This interconnectedness, in conjunction with narrative devices such as cliff-hangers and cutting-off techniques, ensures that earlier episodes increase the appeal and valuation of later ones, creating a stronger incentive for continuous consumption (Linkis, 2021). Third, the value of these complementarities diminishes over time, aligning with the notion that the urge to consume subsequent episodes weakens as consumption capital dissipates over time (Stigler and Becker, 1977;

Heather and Vuchinich, 2003). This time-sensitive aspect of valuation underscores the importance of timing in release strategies and the phenomenon of binge consumption (Schweidel and Moe, 2016; Lu et al., 2019, 2023; Zhao et al., 2022), where the desire to maintain continuity leads consumers to consume episodes in close succession.

Against this backdrop, a personalized monetization strategy that leverages the trade-off between time and money has grown increasingly popular. In particular, firms are beginning to adopt *versioning through wait-times* or "wait-for-free" (WFF) as a business model. Under the WFF policy, consumers make consumption decisions for each episode in an à la carte fashion: they can access an episode immediately for a fixed price or for free after waiting for a pre-specified wait-time. By taking advantage of the complementarity properties of serialized information goods, the platform uses time as a discrimination device to induce consumers with high willingness-to-pay (WTP) to pay for immediate consumption, while preserving the opportunity of free but delayed consumption for consumers with lower WTP. The WFF policy has been embraced by platforms across domains, including Webtoon (comics), Radish Fiction (books), ReelShort (videos) and Real Racing (games).

The goal of this paper is to study how changing wait-times affect consumption decisions and monetization of serialized information goods. Existing research on freemium business models explore various mechanisms that affect consumption and revenues, including advertising revenues (Chiou and Tucker, 2013; Lambrecht and Misra, 2017), product diffusion through referrals (Lee et al., 2019; Kamada and Öry, 2020), network externalities (Shi et al., 2019) and consumer learning (Li et al., 2019; Deng et al., 2022). At the crux of these studies is the trade-off between new user acquisition and cannibalization. Increasing the value of the free option dissuades existing users from paying for the premium version, but at the same time attracts new users who may choose to purchase ("extensive margin"). For serialized information goods, however, its complementarity properties lead to a richer set of consumer dynamics resulting in unique mechanisms that have a strong impact on monetization.

In particular, we introduce the role of *intensive margin*, a mechanism that has not been proposed or examined in prior research. Adapting the terminology from labor and trade economics, we use intensive margin to capture how much a consumer is monetized for a given series, as opposed to extensive margin that captures how many consumers consume the series. Making it easier to wait for free may prompt a consumer to switch from purchasing to waiting for an episode, thereby cannibalizing revenues. However, it may also prompt her to switch from no consumption to waiting, in which case her WTP for the next episode increases due to complementarity leading to positive across-episode spillovers within consumer. By increasing (lengthening) consumption within a series, the firm is able to retain consumers over a longer period and "harvest the acquisition" at times when they receive a positive shock on subsequent episodes.

Moreover, the well-understood aspect of extensive margin is still in play, as more consumers are incentivized to try the series and decide to consume episodes in the series. We aim to provide empirical evidence of these mechanisms and their ensuing impact on aggregate revenues.

Our empirical setting involves a major U.S.-based serialized fiction novel platform that offers series under the WFF policy. We use a rich panel data of over a million users and 10,000 series that covers every consumption record of all users on the platform over a 15-month period, ranging from October 2020 to December 2021. The data set includes whether the user consumed an episode via waiting or purchasing, as well as the wait-time required for a series on any given day. In addition, we augment the data with metadata on series and episodes, such as publication date and promotional activities.

We identify the causal effects of changes in wait-time on user consumption and purchase behaviors by exploiting a natural experiment where the platform unilaterally changed the wait-times for select series during our sample period. In order to increase user engagement, the platform reduced the wait-times for certain series without prior announcement, making it easier for new and existing readers alike to consume episodes for free by waiting. The difference in consumption behavior and aggregate revenues before and after the reduction reflects the effect of wait-time reduction. We estimate the causal effect via a difference-in-differences (DiD) approach by comparing outcome variables of interest within a tight time frame around the reduction.

Two empirical challenges remain: the series for which wait-times were reduced were decided by the platform, which may lead to selection bias, and the reduction was implemented in a staggered manner, which may cause a standard two-way fixed effects model to yield biased estimates. We first address selection bias using a panel matching approach. By constructing a control group of series whose wait-times remain unchanged but have comparable propensity of being reduced, we allow the treatment to be random conditional on covariates, satisfying conditional ignorability. We then mitigate potential concerns around staggered adoption by using a stacked DiD approach. By stacking the panel data based on relative periods around the reduction timing and saturating the model with fixed effects, we can retrieve an unbiased estimate of the causal effect.

Our empirical analysis proceeds in four main parts. First, to explore the impact of changing wait-times on the intensive margin, we examine the total episodes read and purchased in a series by a consumer. By comparing the readers who started the series within 30 days before and after the reduction, we document a XX% increase in the total number of episodes read and a XX% elasticity of consumption. The key estimate of interest is the impact on total episodes purchased, as the increased consumption would only hurt platform revenues if it came at the expense of consumers making less purchases. Our results report no significant change in the total number of episodes purchased, indicating that the cannibalization effect from lower incentives to purchase each episode is offset by the incremental opportunities to make purchases. Moreover, by exploring heterogeneous treatment effects based on platform spending, we find that historical spenders in fact purchase more episodes of the series under reduced wait-times. Thus, we highlight the weakly positive increase in intensive margins – the same consumer who switches from purchasing to waiting for an episode may actually be monetized more in aggregate.

Second, we measure the change in consumption pace in response to reduced wait-times. Given the time-dependent complementarity, shorter wait-times may endogenously increase the consumer's incentives to access the waited episode shortly after the wait-time elapses. By comparing waited consumption during the 30 days before and after the reduction, we find a XX% decrease in the time consumers waited in excess of what is required. This acceleration effect allows the consumers to progress through the series at a faster rate, leading to quicker decisions to purchase subsequent episodes. Third, we explore the impact of reduced wait-times on the extensive margin. We find that the daily count of new readers who consume the series for the first time increases by XX% under reduced wait-times.

Finally, we measure the overall impact of reduced wait-times on daily aggregate consumption and revenues. We find that the reduction on average leads to a 115% and 23% increase in waited and purchased consumption, respectively. This shows that despite lowering incentives to purchase each episode, the shorter wait-times actually uplift platform revenues through both expanded intensive and extensive margins. We then conduct a battery of robustness checks, including Heckman selection model to test for unobservable differences in treated and control series, falsification tests using pseudo treatment indicators and dates, test for potential spillover effects and estimating the model with different model specifications and subsets of the data. The robustness checks indicate broad agreement with our main results.

The rest of the article is organized as follows. First, we discuss how this research is related to previous literature. Second, we describe institutional details and data. Third, we describe the empirical strategy and results and discuss the key findings. Finally, we conclude and provide future research directions.

2 Relationship to the Literature

Our paper contributes to the well-established literature on versioning. Versioning is a widely practiced and studied price discrimination strategy where the firm offers vertically differentiated products at different prices such that consumers with heterogeneous preferences self-select in to the the version-price pair that is targeted to them. Building on the theoretical works on product differentiation (Mussa and Rosen, 1978; Deneckere and McAfee, 1996), Shapiro and Varian (1998) and Varian (2000) laid the groundwork for versioning information goods, pointing out the economic feasibility of manipulating product quality at negligible marginal production costs.

A set of studies have investigated the economic viability of versioning information goods with an emphasis on freemium strategy. Kamada and Öry (2020) models consumers' referral behaviors to show that referral rewards and freemium contracts can be adopted to encourage word-of-mouth. Shi et al. (2019) shows that in the absence of such diffusion dynamics, the freemium model can be optimal when the two products provide asymmetric network externalities. Existing empirical research probes into the impact of introducing the free version on demand for the premium version (Gu et al., 2018; Li et al., 2019; Deng et al., 2022), when firms should charge for content (Lambrecht and Misra, 2017) and how much should be provided for free (Lee et al., 2019; Yoganarasimhan et al., 2022). In many of these cases, the key determinant of the firm's decision is the trade-off between extensive margin and cannibalization: an attractive free offering expands the consumer base, but does so at the cost of cannibalizing existing paid consumption. Our study enriches this dynamic by focusing on the intensive margin. Complementarity between episodes lead to increased consumption in a series per consumer, and the additional purchases made on subsequent episodes offset or even dominate the negative cannibalization effect, leading to greater monetization per consumer.

Moreover, there is limited research that exploits exogenous changes in version quality to empirically investigate the causal implications. Product quality is hard to quantify and empirical settings involving a discrete change in version quality are uncommon. An exception is Li et al. (2019), where they exogenously vary the resolution of free ebook samples. The authors distinguish between "sample quality" and "functional equivalence," where the former is the degree to which the free sample reveals the quality of the premium product (e.g., textual content, image resolution) and the latter is the degree to which the utility derived from the premium product can be obtained from the free product (e.g., limited usage time, hardcover vs. softcover). By varying sample quality, the authors show that under low functional equivalence, it may be profitable for the firm to provide high quality free samples as they serve as poor substitutes. In our study, we exploit an exogenous change in functional equivalence – consumers get full access to the exact same episode by waiting but receive lower utility from delayed consumption. Our results demonstrate that owing to the unique features of serialized information goods, closing the gap between the two versions in terms of functional equivalence can increase firm revenues.

Our work also relates to the literature on sequential product release or using time as a discrimination device. Firms often start with limited distribution through their primary channel and after some time release a secondary channel for mass distribution that sell at a lower margin (e.g., movie theater vs. DVDs, hardcover vs. softcover). If the inter-release timing is too short, forward-looking consumers might hold off on their purchases through the first channel, and if too long, consumers gradually lose interest (buzz decay) and decide to exit the market by the time the second channel opens (Luan and Sudhir, 2022; Calzada and Valletti, 2012; August et al., 2015). While the WFF policy shares commonalities, the release timing is personalized based on the user's consumption timing and is applied to a series of products, giving way to distinct consumer dynamics.

With the proliferation of serialized media content, a growing stream of literature is giving attention to their monetization strategies. Ascarza et al. (2020) highlights the role of customer retention on revenues for mobile games. Using a field experiment where the difficulty of a mobile game is randomized, the authors argue that making it easier for users progress to the next stage increases retention, which leads to higher advertising revenues and long-term user spending. We propose a similar mechanism, whereby shorter wait-times increase the consumers' WTP for subsequent episodes consequently long-term aggregate purchases, while highlighting the unique attributes of serialized information goods. Using data from a comics platform that allows early access for a fee, Choi et al. (2023) finds that habit formation gradually increases consumers' valuation and prompts them to pay for early access. To our knowledge, there are no papers that study the WFF policy with the exception of Choi et al. (2022) due to its novelty. The paper explores a setting where a comics platform adopted WFF policy on a subset of its comics for the first time and finds that the introduction resulted in a boost in free and paid viewership. Our work complements these results by exploiting an exogenous variation in wait-times to identify the causal effect on consumption dynamics and aggregate revenues.

Finally, our work contributes to a nascent stream of empirical research on digital media consumption. Several works provide evidence of binge consumption observed across different product types (Schweidel and Moe, 2016; Lu et al., 2019, 2023). The authors study the impact of binge consumption on downstream behaviors such as responsiveness to advertisements, series completion and spillovers to other content on the platform. Zhang et al. (2022) finds time-inconsistent preferences of readers on a digital book platform, where readers intentionally overpay for content by purchasing individual episodes rather than enrolling in the subscription plan. The authors show that consumers intentionally do so in order to curb future consumption (strategic self-control) and eliminating the pay-per-episode can hurt both consumer welfare and platform profits. Zhao et al. (2022) incorporates the effect of episode release timing on binge consumption, rationing and additional platform visits in the consumer utility model to study the platform's optimal release schedule. The authors find that a hybrid strategy of simultaneous and sequential release strategies yield highest platform profits. We add to this stream of literature by shedding light on the causal link between wait-times and consumption of serialized information goods.

3 Institutional Details and Data

In this section, we describe the institutional details of our empirical setting and subsequently explain the data used in our empirical analysis.

3.1 Institutional Details

The serial fiction market consists of three players: authors, readers and the two-sided platform. Independent authors publish their series that are comprised of multiple episodes on the platform, and the readers access each episode through a mobile application following various payment schemes. The serial fiction market has seen rapid growth globally, 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 that hosts over 10,000 series and has over a million active users.

The platform generates revenues through users' episode purchases. Specifically, each series belongs to one of three sales types depending on the payment scheme: free, premium and "wait-for-free" (WFF). Free series allow users to access all episodes at no cost. Premium series follow a pay-per-episode model where first several episodes are free, and users must pay to "unlock" each subsequent episode. The only difference between WFF and premium series is that WFF allows users to unlock an episode for free once a pre-specified wait-time has elapsed after the last episode of the same series was unlocked. Readers may alternatively use an in-app currency ("Coins") to unlock the episode immediately. Coins can be bought with real money, and every episode costs 3 Coins, roughly equivalent to 50 cents. Users can earn Coins through alternative ways such as watching ads or inviting friends, but they take up a negligible portion compared to direct purchases.

The wait-time varies across series, ranging from 1 to 72 hours, and the same wait-time applies to all episodes and readers within a series. If the consumer becomes eligible to unlock an episode after the wait-time has elapsed, she must actively unlock an episode to "reset the clock" for the next free episode. Take for example a series that requires a 3-hour wait-time. A user may consume the entire series for free as long as she is willing to wait at least three hours *between each episode*. The consumer returning in 12 hours will only have a single free episode available rather than four. Hence, the frequency of visits matters, and one cannot "wait-and-binge", a behavior often observed for series with fixed release schedules. Also, note that there is no incentive for a user to purchase now to read later (i.e., stockpiling), since it will eventually become free after waiting. To be clear, although firms have previously discriminated using time in contexts such as hardcover vs. softcover books, the release timing in WFF is personalized based on the user's consumption of the previous episode and is applied separately for each episode.

Figure 1 illustrates the distribution of series and consumption across the three sales types. Although the distribution of series is roughly even across the three types, consumption for WFF series represent more than 90% of all episodes read on the platform. Given this pattern and our research objective, we focus only on the WFF series within the data.

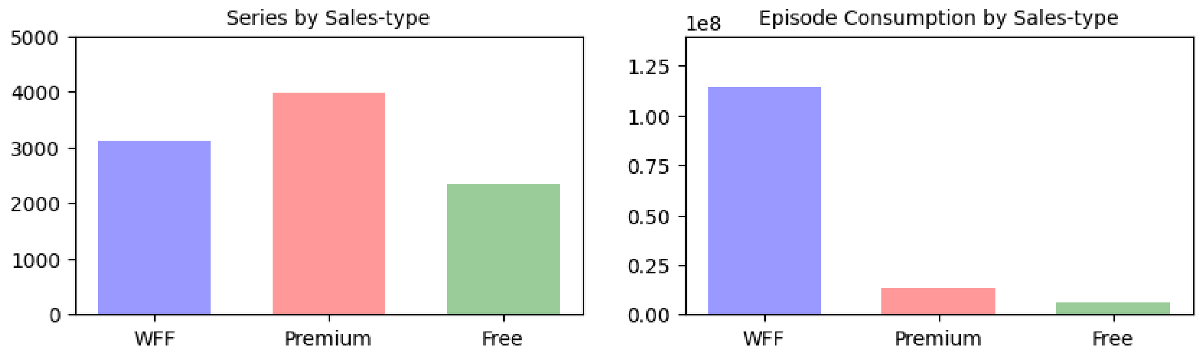

Figure 1: Distribution of series and episode consumption across sales types

Figure 2 illustrates the user experience on the app. The user can scroll through series available on the platform, and once she clicks on a series, additional relevant information is displayed, such as the wait-time, genre and a short description. In this example, "God of Wolves" is a paranormal romance series that requires a 3-hour wait-time and offers the first five episodes for free. The hourglass icon and "3 Coins" indicate that beginning with the sixth episode, the user may either wait 3 hours or pay 3 Coins to unlock.

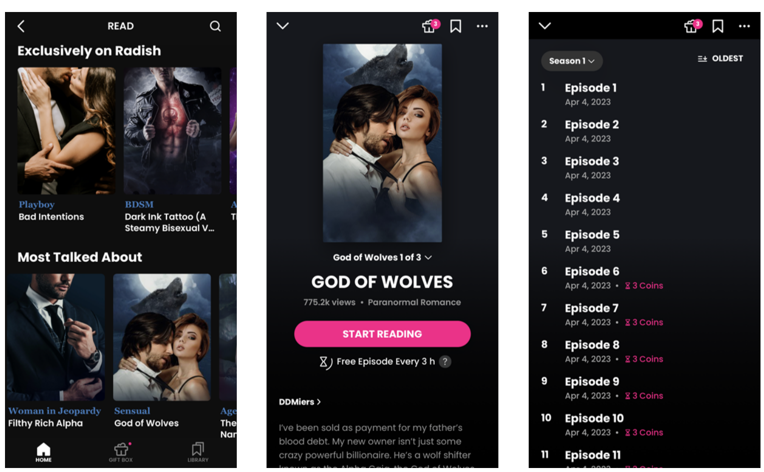

Figure 2: App User Interface

On the left panel of Figure 3 is a sample episode. A typical episode is around 1,500 words, and the vast majority of the readers finish an episode within 15 minutes. At the end of the last unlocked episode, the user is presented with an option to pay to read now or wait to read for free, as shown in the right panel. The pink text on the bottom shows the wait clock ticking down.

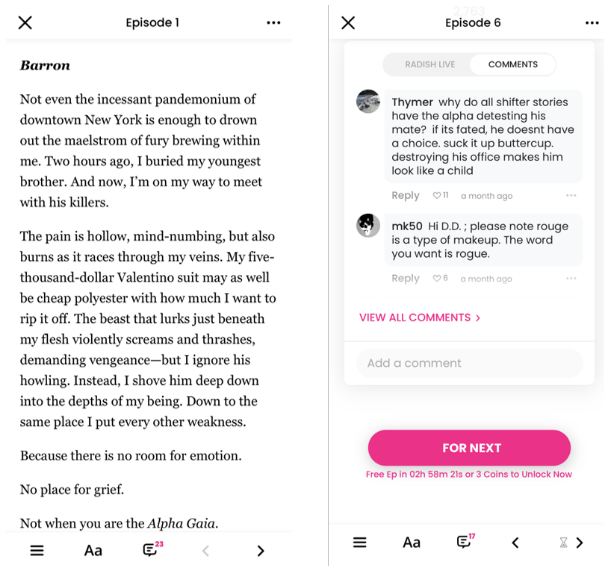

Figure 3: Wait-clock starts ticking down upon accessing the last available episode

3.2 Data

We leverage multiple datasets that cover series and episode metadata as well as user consumption. Series metadata include title, genre, author, sales type and the date of first publication. Episode metadata include series ID, sequence in the series, whether the episode is free, episode publication date and word count. The consumption panel data, which covers 15 months from October 1, 2020 to December 31, 2021, matches every user with every episode that she has read, including how the episode was unlocked: "free" if the episode was free; "waited" if she waited for the required wait-time to elapse; "purchased" if she paid to unlock; "gifted" if she used coupons gifted by the platform (the platform occasionally gifts coupons for specific series that can be used to unlock episodes). We also use series-day level panel data that indicates the required wait-time for each episode and indication of promotional activities such as banner displays or pop-ups on the app.

To isolate the effect of wait-time reduction on the *existing reader base* of the app, we filter the panel data to the readers that joined the platform before October 1, 2020. Moreover, to reduce noise from tail end series that are rarely read, we filter for series with at least 1,000 episodes read during over the entire observation period. Our resulting dataset covers 1,940 WFF series and 308,681 users. Table 1 summarizes basic aggregate information. A median series contains 44 episodes from a single season, and a median user has read two series and 44 episodes during our observation period.

We next provide a set of descriptive statistics. The left panel of Figure 4 illustrates the distribution of series by the size of their reader base, with the x-axis indicating unique reader count (log-transformed) and the y-axis indicating the number of series. The normally distributed histogram shows a heavy concentration of readers on the most popular series. The right panel of the figure illustrates the distribution of users by their

	Mean	SD	Median
Episodes per series	79.1	163.2	44.0
Seasons per series	2.2	4.1	1.0
Series read per user	13.9	39.4	2.0
Episodes read per user	364.0	1003.2	44.0
Episodes waited per user	311.5	1221.7	18.0
Episodes purchased per user	130.4	382.8	16.0
Episodes read per user per day	8.3	10.6	4.0

Table 1: Summary statistics for the main dataset

purchase propensities. Users are heterogeneous in their tendency to purchase versus wait with a bimodal distribution of those that rarely and very frequently purchase. Figure 5 shows the purchase probability (ratio of purchased to total consumption) across episodes for four randomly selected series. Unsurprisingly, there is variation within each series since episode content varies and readers have differentiated tastes.

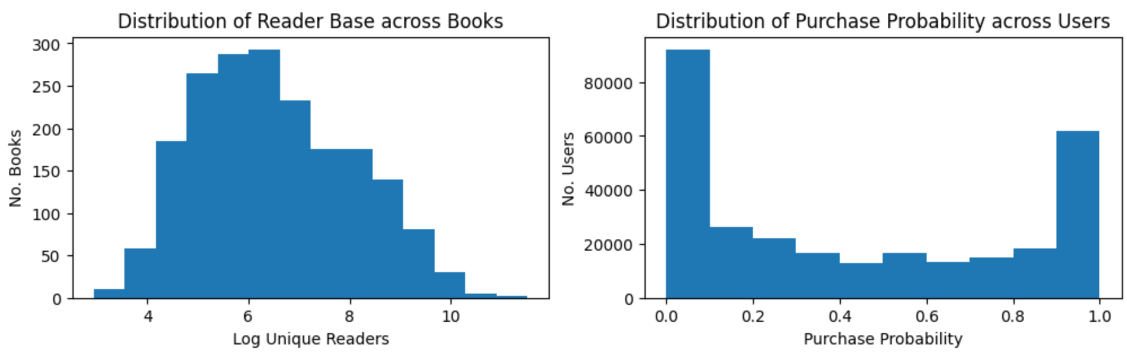

Figure 4: Distribution of readers across series and purchase propensity

As with many serialized information goods such as mobile games, there is very high churn early on. Figure 6 plots the number of readers across the first 20 episodes for the 100 most popular series, normalized by the number of readers for the first episode. The red line indicates the mean across all series, and the blue line is the mean for a subsample with wait-times of at least 12 hours. We find that only about half of the readers that start the series make it past the twentieth episode, less than the halfway mark considering the length of a typical series from Table 1. Moreover, the retention is markedly lower for series with longer wait-times. Next, we sample the consumption panel data for 1,000 randomly selected users. We find that 99% of the episodes are read and along with the immediately preceding episode, and 90% of the episodes are read in sequential order. In other words, the vast majority of readers read the series in the specified order of episodes and read an episode only if they have read the preceding episode. Together with the gradually declining retention plot in Figure 6, this provides patterns consistent with the unique features of serialized information goods: the defined sequence of consumption and the directed complementarity between episodes

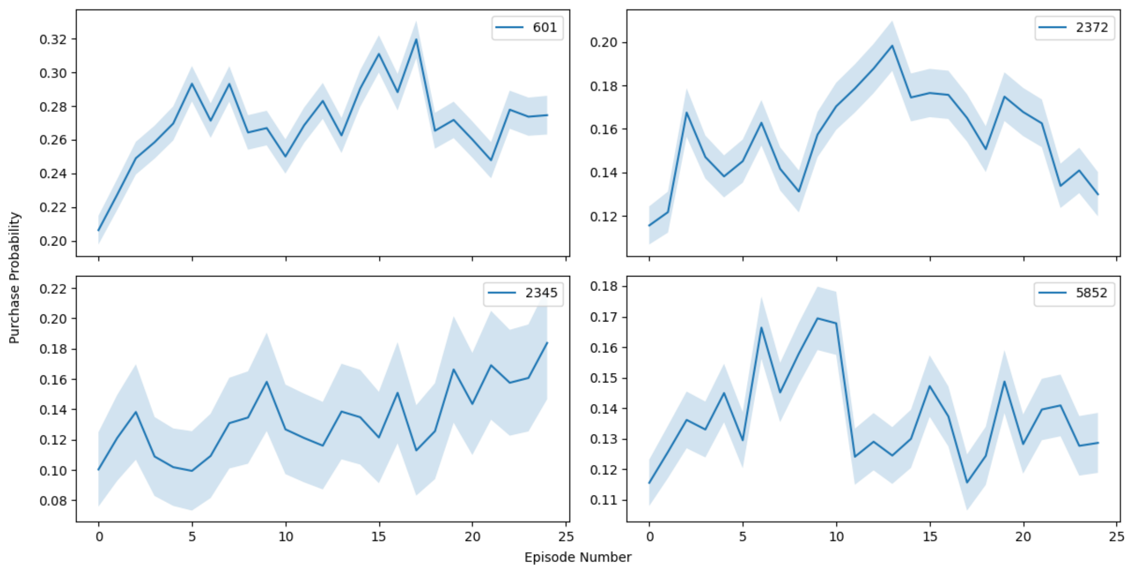

Figure 5: Purchase probability across episodes in a series

that diminishes over time.

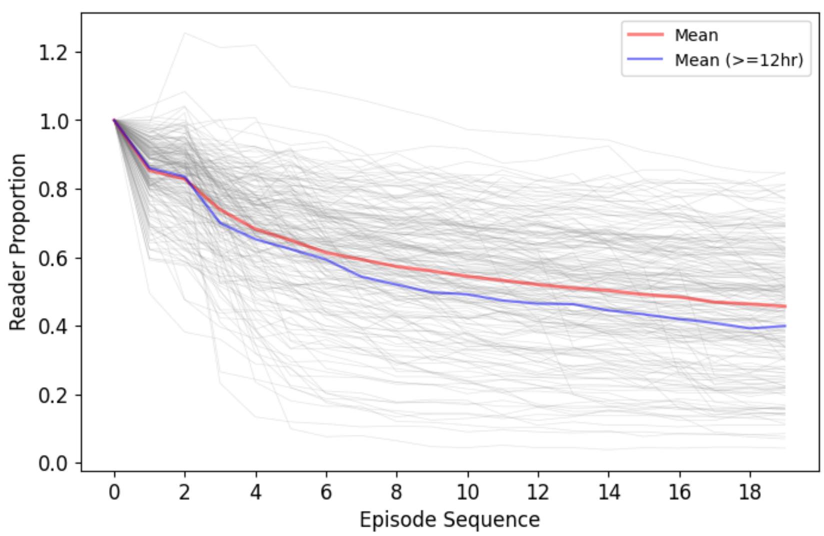

Figure 6: User retention across episodes

We conclude this section by exploring consumption patterns across reading sessions. Because the dataset captures when and for how long each user reads an episode, we are able to define reading sessions at a granular level. For each user, we define a reading session to be one where the interval between finishing an episode and starting the next episode is less than 30 minutes. Table 2 shows the proportion of reading sessions by the number of episodes read and the number of unique series read during the session. Given that 60% of reading sessions consist of a single episode and a median user reads four episodes per day (Table 1), we can infer that users visit the platform multiple times throughout the day mostly just to access the waited episodes and occasionally end up purchasing another episode. Moreover, the table also suggests patterns of binge consumption. 40% of reading sessions involve two or more episodes, and irrespective of session length, at least 25% of sessions involved reading episodes from a single series.

Session Length (eps)	**% Sessions**	**mean**	**25%**	**75%**
1	58.6	1.0	1	1
2	16.2	1.5	1	2
3	7.9	1.8	1	3
4	4.6	2.0	1	3
5	2.9	2.2	1	3
6	2.0	2.3	1	3
7	1.5	2.4	1	3
8	1.1	2.4	1	3
9	0.9	2.5	1	3
10+	4.3	2.7	1	3

Table 2: Unique stories read, by number of episodes in the session

4 Empirical Strategy and Model

4.1 Empirical Strategy

In this section, we discuss our empirical strategy to identify the causal effect of wait-time reduction on reader consumption. An ideal experiment would randomly assign readers to different wait-times for a given series and compare outcomes holding everything else constant. However, our setting only allows the same wait-time for all readers for a given series. Thus, we instead leverage exogenous policy changes implemented by the platform. Specifically, the platform unilaterally reduced wait-times for a certain set of series in varying degrees in a staggered manner within our observation period. For example, users that had to wait 24 hours to unlock an episode would now be able to unlock an episode every hour after the reduction. The platform indicated that the objective of the policy change was to increase overall reader engagement, but the process was rather informal without analytical evidence that called for the change or a systematic criteria. The wait-times before and after the reduction, as well as its timing varies across series, but importantly, there were no prior announcements by the platform regarding the reduction. Hence, readers could not have expected any changes to the wait-time in advance, making it exogenous to them. Similar instances can be found on other platforms where they unexpectedly implemented changes on the terms of the WFF policy such as raising the price of purchased episodes, extending the wait-time for free episodes, or making the waited episode accessible only for a limited duration (Webtoon 2022; Tapas Forum 2022).

We identified 191 series that had wait-times reduced in our dataset, which we call the *treated series*. The rest of the 1,749 series did not have any changes to their wait-times, which we call *non-treated series*. Table 3 lists the number of series from our dataset based on pre- and post-wait-time changes. The diagonal figures represent the series that did not experience changes, and the off-diagonal figures represent the treated series. Note that the treated series are all below the diagonal since the wait-times were reduced. Figure 7 is a graphical representation showing the proportion of series by wait-time before and after the reduction. Figure 8 illustrates the variation in treatment timing and magnitude.

pre/post	1	2	3	4	5	6	7	8	10	12	24	48	72	All
1	1304	-	-	-	-	-	-	-	-	-	-	-	-	1304
2	-	40	-	-	-	-	-	-	-	-	-	-	-	40
3	12	-	20	-	-	-	-	-	-	-	-	-	-	32
4	6	-	-	38	-	-	-	-	-	-	-	-	-	44
5	1	-	-	-	4	-	-	-	-	-	-	-	-	5
6	6	-	-	-	-	7	-	-	-	-	-	-	-	13
7	1	-	-	-	-	-	1	-	-	-	-	-	-	2
8	1	-	-	-	-	-	-	6	-	-	-	-	-	7
10	1	-	-	-	-	-	-	-	1	-	-	-	-	2
12	45	-	3	1	-	-	-	-	-	139	-	-	-	188
24	62	-	21	-	-	-	-	2	-	8	165	-	-	258
48	16	-	-	-	-	-	-	-	-	1	3	23	-	43
72	1	-	-	-	-	-	-	-	-	-	-	-	1	2
All	1456	40	44	39	4	7	1	8	1	148	168	23	1	1940

Table 3: Number of series by wait-time (hrs) for pre- and post- change

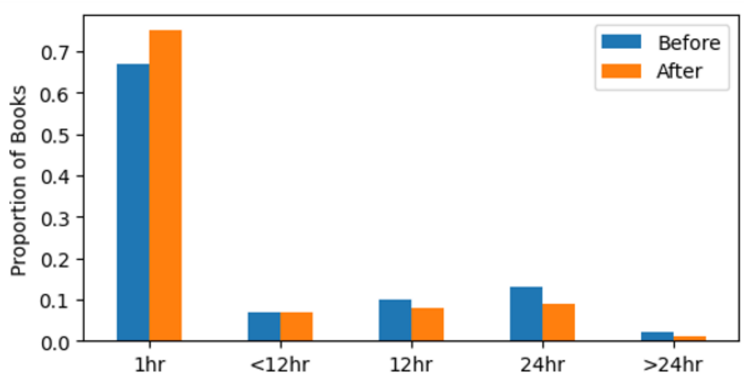

Figure 7: Distribution of wait-times before and after treatment

To identify the effect of shortening the wait-time on individual consumption and aggregate demand, we focus our analysis on a tight window around the reduction. The assumption is that any changes within this brief time period can be attributed only to the wait-time reduction, controlling for a comprehensive set of features. Comparing the treated series to a set of appropriate control series with no wait-time changes, we

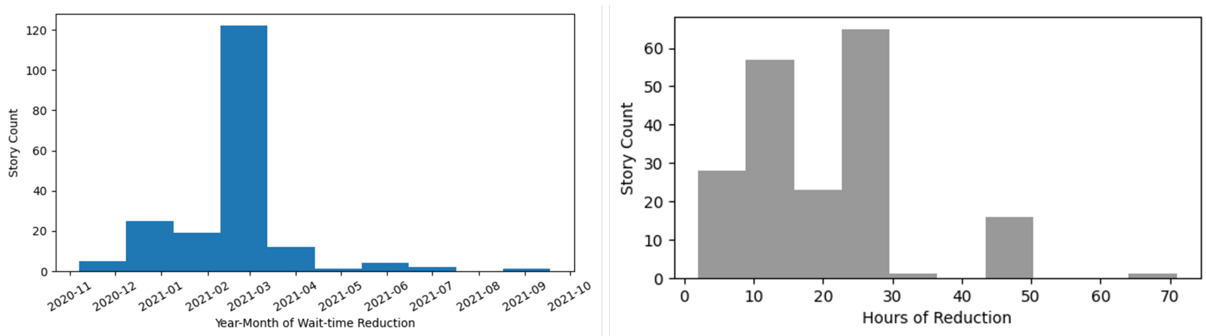

Figure 8: Distribution of treated series by treatment timing and reduction magnitude

can estimate the average effect of wait-time reduction using a difference-in-difference (DiD) framework.

Our empirical context poses two main challenges. The first challenge is the selection into treatment. Although the platform indicated that they did not have a specific selection criteria for the treated series, the selection of treated series is endogenous, which can potentially lead to biased results when naively comparing series that are systematically different. For example, if the treated series have previously been more widely read, then the estimate will be downward biased, as the reduction will have no effect on users that have already read the series. If the platform selected series that were recently updated with new episodes, then the estimate will be upward biased, as it is confounded with by the renewed attention from updated episodes.

The second challenge is that we haven an unbalanced panel data with variation in treatment timing. Since series are published or deleted on the platform at different points in time, the observed time window varies across series. The missing observations can lead to differences in trends before treatment, making the parallel trends assumption difficult to assess and justify. Moreover, recent econometrics literature has shown that variation in treatment timing can lead to biased average treatment effect (ATE) estimates in a two-way fixed effects (TWFE) model, especially in the presence of heterogeneous treatment effects (Borusyak and Jaravel, 2018; de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021). Specifically, the "forbidden comparison" of later treated units to already treated units may assign negative weights to certain sample treatment effects, thereby making the estimated ATE markedly different from the rest of the sample treatment effects.

We address these challenges by using panel-matching approach (Imai et al., 2021) and a stacked DiD model (Cengiz et al., 2019; Deshpande and Li, 2019; Baker et al., 2022; Deng et al., 2022). We first match each of the 191 treated series to a *matched control set* that consists of non-treated series that are fully observed around the treatment timing and have similar probability of treatment. Hence, the treatment is randomly assigned across series in the two groups conditional on observed baseline characteristics, satisfying the conditional independence assumption. We then sample a fixed window around the corresponding treatment dates for each treated series and its matched control set and stack them into a single dataset. This stacked dataset is used to compute an average treatment effect.

4.2 Constructing a Matched Control Set via Panel Matching

We first explore whether the treated and non-treated series are indeed systematically different prior to the reduction. Table 4 presents time-varying features across the two groups, as well as the p-values to compare the distributions. We observe the average daily number of episodes read by waiting, purchased, and unique readers, as well as binary indicators for gifted episodes, promotions and new episodes published. For the treated series, features are measured over the four weeks before their respective treatment dates; for the non-treated series, features are measured over the four weeks before the earliest treatment date (or the first four weeks of publication if the former time-frame is less than four weeks). We see that the two groups are significantly different across almost all features, which suggests a simple comparison between the two groups can lead to confounding and motivates the need for a matched control group to satisfy the conditional independence assumption, $(Y(0), Y(0)) \perp\!\!\!\perp T|X$.

	Untreated	Treated	p-value
$\log(waited + 1)$	3.54	3.46	0.506
$\log(purchased + 1)$	2.68	2.92	0.027
$\log(readers + 1)$	2.92	3.39	0.000
$\mathbf{1}(gifted > 0)$	0.11	0.06	0.075
$\mathbf{1}(promo)$	0.09	0.02	0.000
$\mathbf{1}(new\ episode)$	0.28	0.19	0.016

Table 4: Comparison of treated and untreated series

In order to address the systematic differences, we create a control group for each treated series by matching it with non-treated series that are similar in terms of observed characteristics prior to treatment. By making treatment independent of observed potential confounders, we are able to draw causal conclusions about the impact of reduced wait-time by comparing the two groups. We utilize the propensity score matching procedure for time-series cross-section data (panel-matching) developed in Imai et al. (2021). Despite the popularity of matching methods, almost all of the existing methods assume a cross sectional dataset using static features measured at a point in time (Abadie and Imbens, 2011; Diamond and Sekhon, 2013; Hansen, 2004). Studies involving a panel dataset compute the average of time-varying covariates over a static time-frame (Datta et al., 2018; Deng et al., 2022; Narang and Shankar, 2019), which can miss out on important time-varying factors that affect selection into treatment.

In our setting however, the potential demand-related confounders (e.g., waited/purchased consumption, number of readers) are time-varying, and the variation in treatment timing makes it difficult to define a pretreatment 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 trajectories. For example, a series that is gaining traction among readers and one that is becoming increasingly unpopular will clearly experience different effects from reduction in wait-time. Furthermore, since we have an unbalanced panel data with staggered treatment adoption, we must match each treated series to non-treated series that are observed in the same time window. Figure 9 is a treatment variation heatmap from a random sample of series. Each row represents a 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 (weeks when the series was not on the platform). We want to match each treated series to non-treated series that are observed (blue areas) around the treatment timing and are comparable in covariate values leading up to treatment.

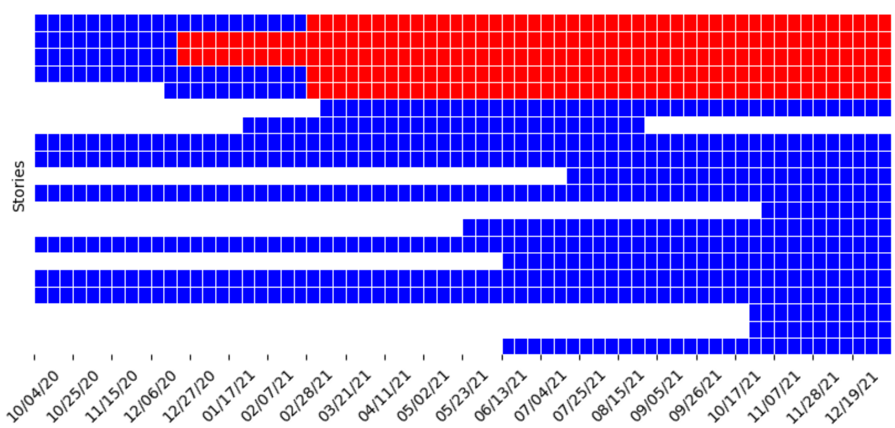

Figure 9: Treatment Variation Plot

Constructing the Matched Control Set We now describe the matching procedure in detail. Let us denote a treated series s that receives treatment in period t as observation (s, t). For each treated observation (s, t), we construct the matched set of never-treated units that are fully observed from time $t - L$ to $t - 1$. Imai et al. (2021) allows the matched set to include not-yet treated units, but we only allow for never treated units for a more robust comparison. Figure 10 illustrates an example of how matched sets are created when $L = 2$. In this example, observation $(0, 2)$ is matched to $(3, 2)$ and $(4, 2)$ as the two are never-treated and are observed in $t = 0, 1$. Observation $(5, 2)$ is not included in the matched set as it is not observed in $t = 0$. In our case, we set $L = 4$, which assumes that adjusting for covariate trends up to four weeks back removes most of the possible confounding. Formally, the matched set 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 and 0 if not.

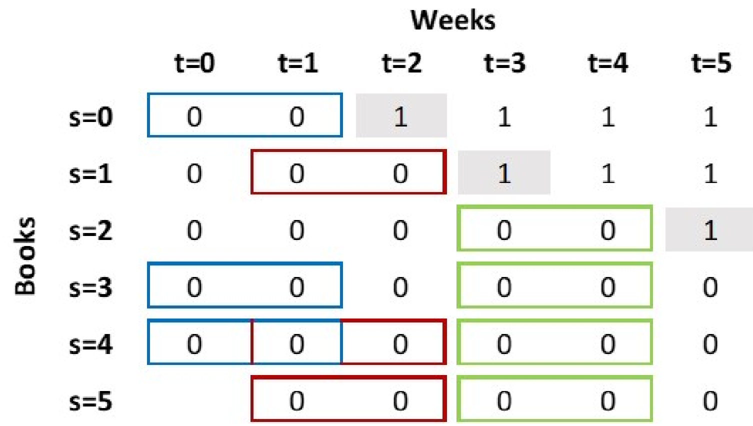

Figure 10: Treatment matching

Refining the Matched Sets The previous matched sets only account for treatment and observation history. Next, we refine the matched sets based on propensity scores, the conditional probability of treatment assignment given observed covariates (Rosenbaum and Rubin, 1983). The propensity score is computed using a logistic regression based on a rich set of observed covariates that can reasonably discriminate the treated and non-treated series:

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

where $X_{s,t}$ is a matrix of observed time-varying covariates for series s in week t. The covariates used in the logistic regression include four weekly waited/purchased consumption, number of unique readers, binary indicators for gifted episodes, as well as single binary indicators for promotion and new published episodes during the four-week period (18 variables in total).

Given the fitted model, we compute the estimated propensity score $\hat{e}_{st}$ for all treated observations and their matched sets. Among the series in the matched set whose propensity score distance to the treated unit is less than a defined caliper ($C = 0.1$), we select up to N series (or all units if fewer than N satisfy the criterion) with replacement. In this way, we choose a subset of the original matched set that are most similar to the treated unit in terms of the observed confounders. Formally, the refined matched set 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 set.

Covariate Balance Diagnostics The number of matches, N, is set such that the best covariate balance between the treated series and their matched counterpart is achieved. Figure 11 compares the quality of covariate balance across $N = 1, \ldots, 10$. On the left panel, the x-axis indicates the number of matches, and the y-axis indicates the number of covariates that are diagnosed to be balanced based on standardized mean differences (SMD) and p-values. Lower SMDs indicate stronger balance, with literature suggesting 0.1-0.15 as the cutoff for indication of good balance (Stuart et al., 2013; Zhang et al., 2019). The plots show good balance for $N \geq 3$. The right panel of Figure 11 plots the Kullback–Leibler divergence (KLD) for the propensity scores between the treated and matched control series by the number of matches. KLD closer to zero indicate greater similarity between two distributions. We see that the similarity between the two groups are stable across N.

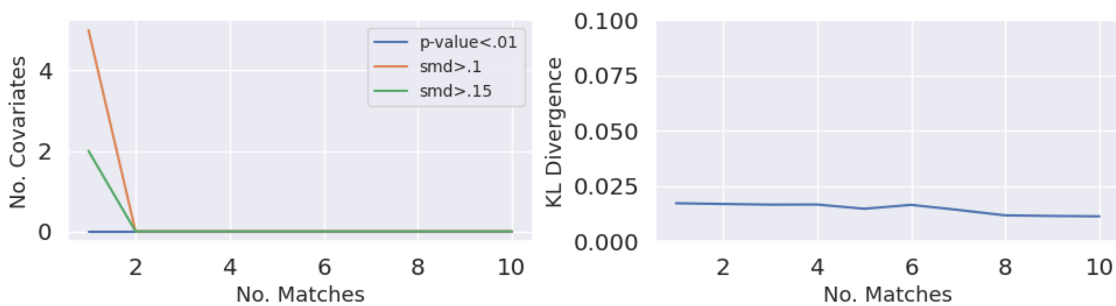

Figure 11: Assessment of match quality by the number of matches

Matching a treated unit to a single or multiple control units are both widely used in practice, each with its tradeoffs. First is the precision-bias tradeoff. One-to-one matching tends to produce less biased estimates because each treated unit is matched with its closest match, but it may yield fewer matches, resulting in less precision. On the other hand, 1:N matching could increase precision by utilizing more available data but may introduce bias if the matches are not as close. Since we set the propensity scores of matched pairs to be less than a capliper $C = 0.1$, potential bias issues are limited. In terms of sample size, 1:N matching can potentially utilize all available control units, thereby maintaining a larger sample size and consequently reducing variance. At the same time, 1:N matching runs a risk of overfitting if one uses too many matches for each treated unit. In order to secure a large enough sample while against potential overfitting, we proceed with $N = 6$, beyond which some treated series start to have less eligible matched series. We conduct robustness checks to show that the results of the analysis remain unchanged for $N \geq 5$.

To evaluate the quality of propensity score matching, we check for covariate balance and overall distributions of propensity scores before and after matching (Caliendo and Kopeinig, 2008; Haviland et al., 2007). The results in Table 5 show good balance across every covariate and significant improvements in SMD before and after matching. Figure 12 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 strong match.

	Means control (before matching)	Means control (after matching)	Means treated	p-value	% Reduction in SMD
Propensity Score	0.399	0.601	0.602	0.916	99.302
$\log(T1\ waited + 1)$	4.989	5.293	5.305	0.917	95.874
$\log(T2\ waited + 1)$	5.003	5.198	5.267	0.578	73.473
$\log(T3\ waited + 1)$	5.000	5.220	5.258	0.761	84.712
$\log(T4\ waited + 1)$	4.987	5.146	5.222	0.562	67.620
$\log(T1\ purchased + 1)$	4.006	4.683	4.594	0.519	83.657
$\log(T2\ purchased + 1)$	4.020	4.728	4.655	0.575	87.119
$\log(T3\ purchased + 1)$	4.015	4.598	4.511	0.545	81.310
$\log(T4\ purchased + 1)$	4.020	4.527	4.455	0.623	82.663
$\log(T1\ readers + 1)$	3.643	4.148	4.177	0.788	94.235
$\log(T2\ readers + 1)$	3.655	4.101	4.141	0.716	91.380
$\log(T3\ readers + 1)$	3.657	4.112	4.131	0.868	95.890
$\log(T4\ readers + 1)$	3.659	4.076	4.096	0.858	95.321
$\mathbf{1}(T1\ gifted{>}0)$	0.044	0.032	0.026	0.824	62.394
$\mathbf{1}(T2\ gifted{>}0)$	0.044	0.028	0.031	0.813	68.909
$\mathbf{1}(T3\ gifted{>}0)$	0.043	0.015	0.016	1.000	95.706
$\mathbf{1}(T4\ gifted{>}0)$	0.045	0.010	0.010	1.000	95.883
$\mathbf{1}(promo)$	0.048	0.012	0.016	0.724	84.002
$\mathbf{1}(new\ episode)$	0.139	0.162	0.188	0.401	48.513

Table 5: Similarity between treated and control groups after matching

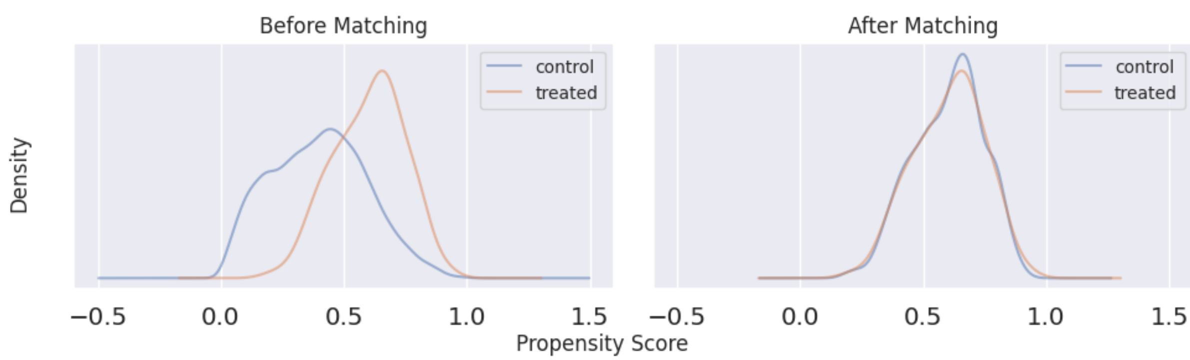

Figure 12: Propensity score distribution before and after matching

4.3 Econometric Model

We now turn to the econometric model to estimate the causal effect of the reduction in wait-times. To control for unobservable time-trends, we estimate the treatment effect in a difference-in-difference framework controlling for a set of fixed effects and covariates. To retrieve an unbiased measure of ATE, we must address the issue of variation in treatment timing, which can be problematic for standard TWFE models. When treatment effects vary across both time and units, then the TWFE estimand of ATE may correspond to a non-convex weighted average of individual treatment effects. Hence, we utilize a stacked difference-in-difference model. Stacked DiD has been suggested and widely used 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; Deng et al., 2022). Gardner (2022) shows that this approach estimates a convex weighted average of the individual treatment effects under parallel trends and no anticipation.

We start by constructing event-specific datasets of equal length for each of the 191 treated series. The dataset includes the outcome and control variables of the treated series and its six matched control series, which we denote as a *series group* consistent with the notation from Deng et al. (2022). Note that although a non-treated 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 TWFE DiD regression with series group-specific series and time-fixed effects, which fully capture the unobservable characteristics. This technique essentially estimates the DiD from each series group and then applies variance weighting to aggregate the treatment effects (Baker et al., 2022).

The stacked DiD model takes the following form:

$$Y_{sgp} = \beta^{DD}(after_p \times treated_s) + X_{sgp}\gamma + \delta_{sg} + \nu_{gp} + \epsilon_{sgp} \tag{4}$$

where s denotes series, g denotes series group, p denotes period, and Y_{sgp} denotes the main dependent variable measured for series s of series group g in period p. $treated_s$ is a binary treatment indicator for series s, and $after_p$ is a post-treatment dummy for period p. The main coefficient of interest is β^{DD}, the average treatment effect of wait-time reduction. X_{sgp} is a matrix of observable control covariates; δ_{sg} is a fixed effect specific to series s in series group g that captures time-invariant unobservable characteristics (referred to as *Series-series group FE*); ν_{gp} is a fixed effect specific to group g in period p, which captures unobservable time trends, allowing for different trends in different groups (referred to as *Series group-period FE*). ϵ_{sgp} is the error term, which are clustered at the series level.

4.3.1 Identifying Assumptions

Causal identification holds under the assumptions of parallel trends, no anticipation and the stable unit treatment value assumption (SUTVA). The parallel trends assumption requires that the treatment group would have had an identical trend to the control group had the treatment not been implemented, and the no anticipation assumption requires that in periods prior to treatment, the outcome variable was not affected by the upcoming treatment. If these two assumptions hold, then any time-varying unobservables will be absorbed by the trends in the control group. For these to hold, we must be confident in the validity of the control group. The platform unilaterally implemented the reduction without notifying the users in advance, which prevents any strategic action from the users such as delaying consumption or purchase. Moreover, the matching process conducted in the previous section ensures covariate balance between the treated and non-treated series, making the treatment plausibly random conditional on the propensity score. This contention is strongly supported by the parallel pre-event trends for the different outcome variables in Figure 13. The figure shows for the two groups, the trend in daily aggregate around the wait-time reduction date. For control series, we first calculate the median across all control series within a series group and then calculate the average value across all series groups. The difference in trends prior to the reduction is insignificant for all outcome variables.

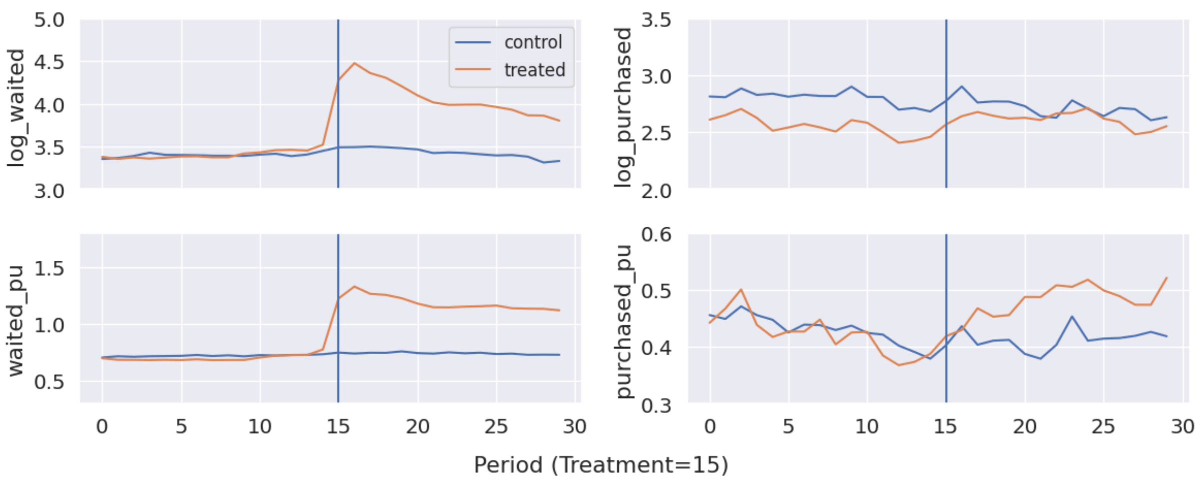

Figure 13: Outcome variable trends before and after treatment

SUTVA assumes that the potential outcomes of each unit are not influenced by the treatment assignment or outcomes of other units (i.e., no spillover effect between units). Under SUTVA, the treatment effect for a given unit is independent of the treatment assignment and outcomes of other units. Our research design makes this assumption plausible. The platform hosts over 10,000 series on the platform (also including free and premium series), and treatment was implemented at various times across only 191 series. Hence, it is unlikely that a reduction in wait-time for any series had a meaningful spillover effect on other series on the platform. Figure 13 also provides descriptive evidence that treatment spillovers are unlikely to have a dominant effect, because there is no discernible change in the trends in the control groups after treatment period. Nonetheless, we later conduct robustness checks to explicitly rule out the possibility of spillovers and show that treatment does not have significant spillover effects on matched control series.

5 Results and Robustness Checks

5.1 Impact of Wait-time Reduction on Consumption

The results of our main analysis presented in Table 6 demonstrate a positive and significant causal effect of the wait-time reduction on free and paid consumption. Column (1) shows the impact on waited reads, $\log(waited + 1)$. Denote the daily waited episodes before the wait-time reduction as N_0. The percentage change in waited episodes from the treatment can be computed using $\hat{\beta}$, the estimated coefficient of $after \times treated$, as: $(e^{\hat{\beta}}(N_0 + 1) - N_0 - 1)/N_0$. The estimate 0.761 in column (1) suggests that if the daily waited episodes before the reduction is at the mean (116), and then if all else equal, it would increase to 250, a 115% increase. This result is expected as shorter wait-times by design allow readers to unlock free episodes more frequently. The *purchased* consumption is the platform's primary concern as it is directly linked to revenues and could be subject to cannibalization as purchasing readers substitute to waiting. It would be troubling for the platform if waited reads increased at the expense of purchased reads. Column (2) shows that those concerns are unwarranted and rather shows the opposite. It suggests that if the daily purchased episodes before the reduction is at the mean (58), it would increase to 71, a 23% increase. Although the percentage increase is smaller than that for waited reads, it is surprising that the shorter wait-times also caused readers to pay for more episodes. In columns (3) and (4), we analyze changes to consumption at the per reader level. We find that consumption per unique reader also significantly increases. If the daily waited episodes per reader is at the mean (1.06), it would increase to 2.47, a 133% increase. The mean of daily purchased episodes per reader increases from 0.65 to 0.80, a 23% increase.

In the dynamic model specification, we estimate separately the effect for each day within the 15 days before and after the wait-time reduction. Figure 14 plots the estimated β coefficient when allowing the effect to vary by day. Consistent with Figure 13, the estimated treatment effects are insignificant prior to treatment, while consistently positive and significant post treatment. It also illustrates the absence of a pretrend prior to treatment, further justifying our DiD specification.

5.2 Robustness Checks

In this section, we conduct a battery of robustness checks and tests to rule out alternative explanations that might explain the increased free and paid consumption from reduced wait-times.

	(1) log(waited + 1)	(2) log(purchased + 1)	(3) waited pr	(4) purchased pr
$after \times treated$	0.761***	0.207***	1.412***	0.147***
	(0.051)	(0.053)	(0.061)	(0.034)
log(*free eps*)	0.095	0.177	0.390	0.262*
	(0.165)	(0.185)	(0.252)	(0.134)
log(*paid eps*)	1.588***	0.823**	0.981***	-0.085
	(0.267)	(0.330)	(0.236)	(0.203)
log(fp)	-0.141	-0.039	-0.505**	0.063
	(0.295)	(0.244)	(0.198)	(0.133)
log(*lp*)	-0.098***	-0.166***	0.130***	0.012
	(0.024)	(0.028)	(0.033)	(0.014)
log($gifted$)	0.088***	0.077***	-0.024**	-0.026***
	(0.016)	(0.016)	(0.012)	(0.010)
log(*T7 gifted*)	0.130***	0.106***	0.056***	0.003
	(0.017)	(0.014)	(0.015)	(0.006)
$promo$	1.525***	1.320***	-0.030	-0.105
	(0.222)	(0.180)	(0.229)	(0.111)
T7 promo	1.919***	1.701***	0.924***	-0.023
	(0.106)	(0.106)	(0.110)	(0.066)
R-squared Adj.	0.211	0.064	0.229	0.002
N Obs	40,080	40,080	40,080	40,080
N Series	636	636	636	636
DOW FE	Y	Y	Y	Y
Series-series group FE	Y	Y	Y	Y
Series group-period FE	Y	Y	Y	Y
Time FE	Y	Y	Y	Y

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 6: Difference-in-Difference Results

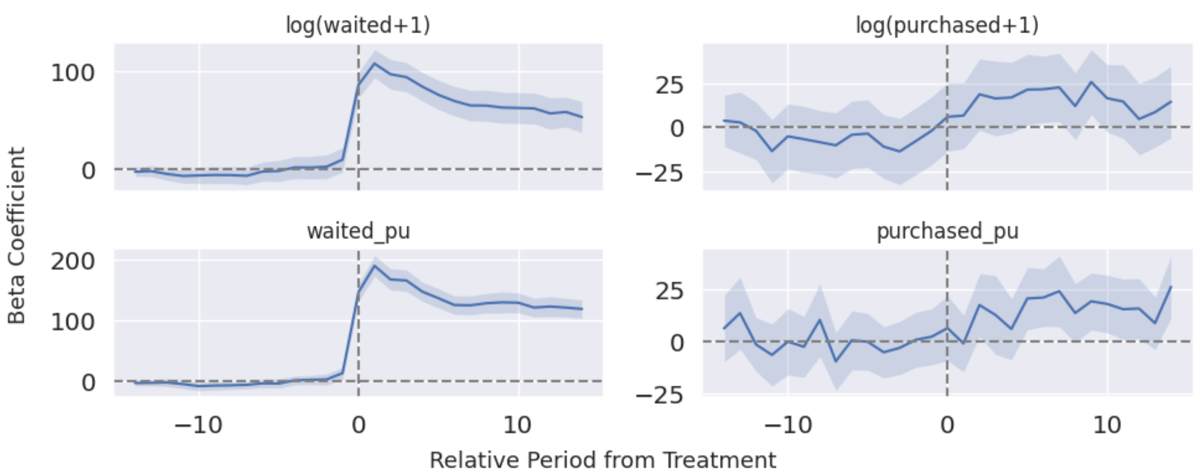

Figure 14: Estimated treatment effect by period with 95% confidence interval

5.2.1 Potential Selection on Unobservables

Although we ensure for our main analysis that the treated and matched control series are similar based on observed characteristics, the two groups may still differ on unobserved factors which may lead to biased results. To formally account for potential unobservable differences, we use the two-stage Heckman correction procedure (Heckman, 1979). The Heckman correction first estimates the probability of a unit being selected into treatment (selection equation) and then uses this information to correct the outcome equation for selection bias.

The Heckman model requires a variable that satisfies the exclusion restriction – there should be at least one variable that affects selection in to treatment but does not directly affect the outcome variable in the second stage. In this context, the variable should influence the platform's decision to reduce wait-times, but not directly the demand for readers' waited and purchased consumption. A potential source of such exogenous variation is the *program type*. Every series belongs to one of seven program types that the platform uses for internal classification purposes: UGC, Minimum Guarantee, Flat Fee, Publisher, Original, Acquired Original and Pilot. UGC (user-generated content) is the most common type, referring to series by independent authors where the revenue proceeds follow a percentage split between the author and the platform. Minimum Guarantee is similar but has a lower limit on the proceeds to the author; Flat Fee has a predetermined payment to the author; Publisher refers to series under contract with a publication company; Original and Acquired Original are series owned by the platform; and Pilot refers to a handful of series yet without a formal contract. Figure 15 shows the distribution of series by program type.

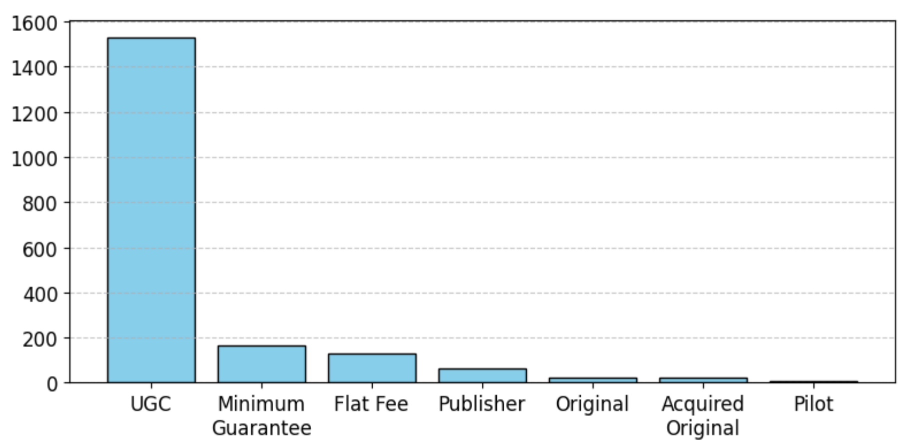

Figure 15: Distribution of series by program type

Of the 191 treated series were 190 UGC and 1 Original series. Presumably, the program type has implications on both the feasibility of adjusting the wait-times and the revenues that accrue towards the platform, which is likely to influence selection in to treatment. To satisfy the exclusion restriction, the program type should not drive consumption through avenues other than wait-time reduction. As it is impossible for readers using the app to know the contract details determined behind the scenes, we can make a convincing case that the program type satisfies the exclusion restriction.

One concern might be that the program type is endogenous – program type may be correlated with the quality of the series, which may directly influence consumption. It could be the case that prior to publication, the platform forms an expectation on how much traction a series will get, either based on the storyline or the author's previous track record, and pushes for the program type most beneficial to the platform. It also could be the case that Original series produced in-house are of higher quality than series by independent writers. If so, the program type may be correlated with consumption and fail to satisfy the exclusion restriction. To alleviate the concern, we examine the correlations between the program type and various proxies of series quality. Since 79% of the WFF series are categorized as UGC, we separate the series into UGC and non-UGC. Then, we compute the correlation against daily average count of episodes waited and purchased measured across the first 60 days observed in the data. The correlations are low at -0.095 and -0.084, respectively. The density plots of waited and paid consumption for the two group of series over the 60 days shown in Figure 16 indicate no significant difference in the proxies for quality level. Thus, it is unlikely that the quality of the series was endogenously driving the program type. Moreover, we control for those proxies of quality in our matching algorithm.

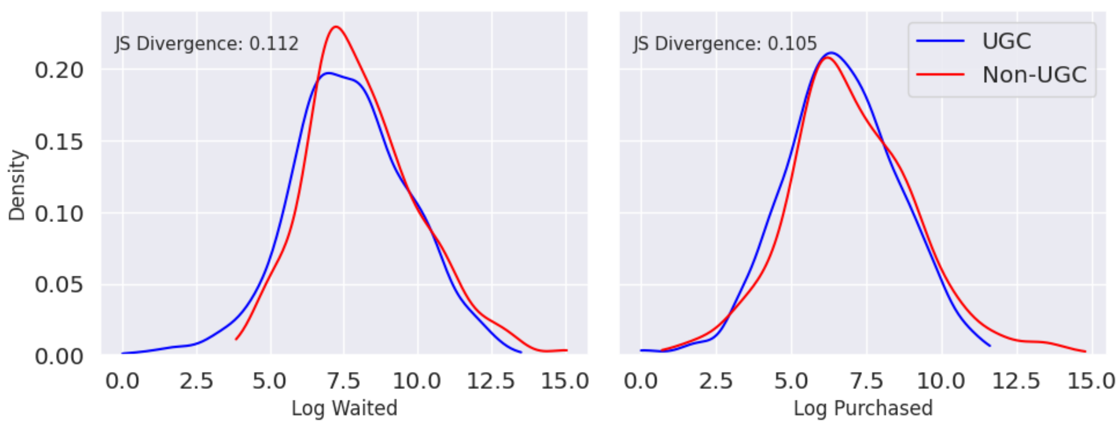

Figure 16: Density plot of free/paid consumption for UGC and non-UGC series

In the first stage, we take the matched dataset resulting from PSM and use a probit model to estimate the probability of selection into treatment as a function of program type and other covariates that affect selection into treatment:

$$Pr(D_s = 1|ProgramType_s, X_s, \epsilon_s) = \Phi(\alpha_0 ProgramType_s + \alpha_1 X_s + \epsilon_s) \tag{5}$$

where $ProgramType_s$ is an indicator equal to 1 if series s is UGC, X_s is the vector of other covariates similar to those used in propensity score matching, and ϵ is the error term. We compute the inverse Mills ratio from this probit regression. We then augment the regression model in Equation 4 by including the inverse Mills ratio as an additional covariate in the second stage. Table 7 reports the results for the first-stage selection model. We see that UGC series are more likely to receive a reduction in wait-times. The results also suggest that the treated series are more deeply read (positive coefficients on $\log(waited + 1)$), but less widely read (negative coefficient on $\log(readers + 1)$). Given that the underlying goal of the platform was to increase readership, it is reasonable that the platform chose series with sufficient appeal conditional on trial, but were not able to acquire a large audience.

Variable	Coefficient (standard error)
$ContractType$	1.42*** (0.36)
$\log(waited + 1)$	0.51*** (0.18)
$\log(purchased + 1)$	-0.04 (0.08)
$\log(readers + 1)$	-0.47** (0.22)
$\mathbf{1}(gifted > 0)$	0.41** (0.21)
$\mathbf{1}(promo)$	-0.20 (0.36)
$\mathbf{1}(new\ episode)$	0.41*** (0.12)
$\mathbf{1}(wait\text{-}time > 12)$	1.19*** (0.17)
Intercept	-3.43*** (0.40)
N	1,336
Pseudo R-squared	0.13

Table 7: First-Stage Probit Model Results

The results of the stacked DiD with the inverse Mills ratio (IMR) as an additional covariate is shown in Table 8. We do not find evidence for selection as the selection correction term IMR is insignificant. Hence, the coefficients on $after \times treated$ largely remain unchanged as expected, indicating robustness of our main findings to potential selection on unobservables.

	(1) log(waited + 1)	(2) log(purchased + 1)	(3) waited pr	(4) purchased pr
$after \times treated$	0.748***	0.202***	1.406***	0.145***
	(0.059)	(0.060)	(0.074)	(0.040)
IMR	0.956	0.438	0.513	-0.302
	(0.788)	(0.475)	(0.337)	(0.455)
R-squared Adj.	0.920	0.810	0.685	0.274
N Obs	40080	40080	40080	40080
N Series	636	636	636	636

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 8: Stacked DiD Results with Inverse Mills Ratio

To further show robustness, we conduct a sensitivity test assessing how robust the DiD results are to the threat of unobserved confounders following Cinelli and Hazlett (2020). By extending the omitted variable bias framework, the authors develop an analysis that provides, relative to observed covariates, how strongly unobserved confounders need to be associated with both the outcome and treatment variables (in terms of partial R^2) to explain away the estimated treatment effect. As the benchmark observed covariate, we do not rely on a single variable, but rather select multiple uncorrelated covariates from Table 6 to be conservative: log(paid eps), log(lp), log(T7 gifted) and log(T7 promo). The results are presented in Figure 17. It shows that even if unobserved confounders explain the variation of residuals in the waited and purchased consumption twice as much as the benchmark covariates *combined*, the estimated treatment effects are positive and significant.

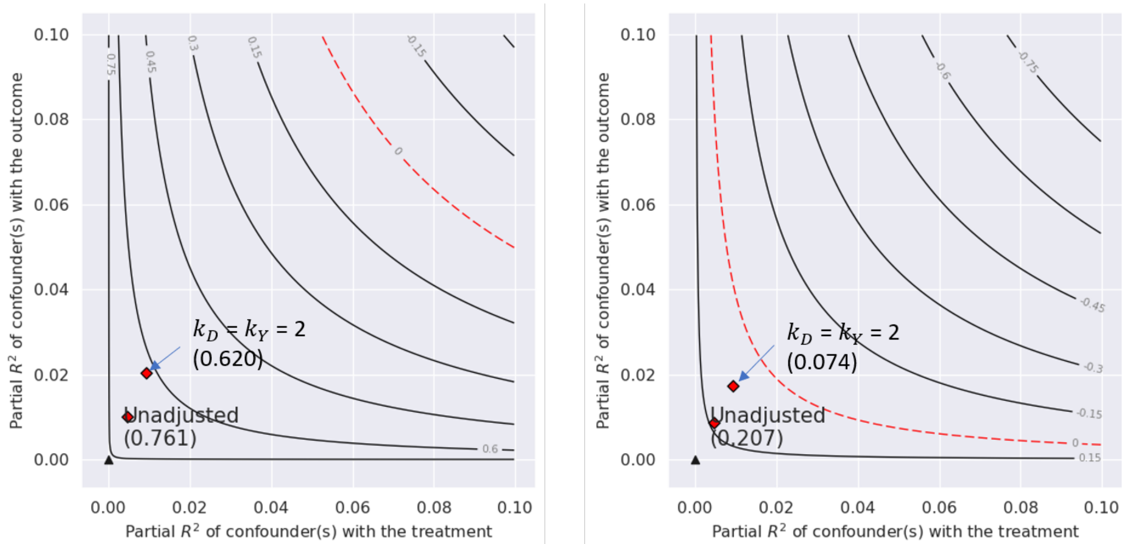

Figure 17: Sensitivity contour plots of estimated treatment effect on waited (left) and purchased consumption (right)

5.2.2 Falsification Tests

We test the possibility that the estimates in Table 6 are coincidentally picking up spurious effects by using pseudo treatment indicators and dates. For pseudo treatment indicators, we randomly assign a control series as treated for each matched series group and estimate the same model without the actual treated series. Under this falsification test, since the pseudo treatment indicator does not reflect the true information of whether the wait-time of the series is reduced, the estimated treatment effects should be insignificant (Ghose and Todri-Adamopoulos, 2016; Jo et al., 2020). For the pseudo treatment dates, we manipulate the treatment date to be 15 days prior to the actual date when the wait-time was reduced. Since the modified timeframe does not include the actual treatment date, the estimates should again be insignificant. Table 9 shows that the coefficients on $after \times treated$ are indeed statistically insignificant for both specifications, indicating that our findings are not a statistical artifact of our specification.

	(1) log(waited + 1)	(2) log(purchased + 1)	(3) waited pr	(4) purchased pr
Pseudo Treatment Indicators				
$after \times treated$	0.069	0.069	0.026	-0.016
	(0.052)	(0.062)	(0.034)	(0.037)
R-squared Adj.	0.200	0.074	0.054	0.001
N Obs	34350	34350	34350	34350
N Series	445	445	445	445
Pseudo Treatment Dates				
$after \times treated$	-0.009	0.006	-0.046*	-0.032
	(0.044)	(0.055)	(0.025)	(0.034)
R-squared Adj.	0.154	0.054	0.045	0.001
N Obs	39570	39570	39570	39570
N Series	630	630	630	630

Note: Robust standard errors clustered at series level in parentheses.

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

Table 9: Results of Falsification Tests using Pseudo Treatment Indicators and Dates

5.2.3 Potential Violation of SUTVA

In order to make a casual claim from our DiD model, we must ensure that there are no significant spillover effects between series from the treatment. For example, there may be potential substitution where readers move away from non-treated series to spend more time reading stories with reduced wait-times. There also may be complementary effects where readers consume episodes of the treated series more frequently and in doing so also read the untreated series more frequently. In Section 4.3.1, we argued that significant spillover effects are unlikely because the platform hosts over 10,000 series (including free and premium series), and only a small fraction of the series received treatment in a staggered manner. Figure 13 also shows no visible change in the control series after treatment.

To provide further evidence, we statistically test for violation of SUTVA using the overlap in reader base between treated and control series. The idea is that if there are any substitution or complementary spillover effects, it should be the non-treated series whose majority of readers also read the treated series that is impacted the most. Specifically, we measure the percentage overlap in readers between each control and treated series in every series group. We define $Overlap_{sg}$ as the proportion of readers of series s who have also read an episode of the treated series in series group g during the 15 days prior to treatment. We include an interaction term $after \times overlap$ in the main model and estimate the coefficients.

$$log(Y_{sgp} + 1) = \beta D_s T_p + \rho Overlap_{sg} T_p + X_{sgp}\gamma + \delta_{sg} + \nu_{gp} + \tau_g + \epsilon_{sgp} \quad (6)$$

A significant coefficient on the interaction term ρ would suggest spillover effects from treatment, positive values indicating complementarity and negative values indicating substitution. Results are shown in Table

10. The main treatment effect remains unchanged, while the coefficient on $after \times overlap$ is not significant, providing evidence for satisfaction of SUTVA. For further robustness, we also allow up to ten matched control series per treated series, which would check potential spillover effects for a wider set of series, and find no qualitative difference.

	(1) log(waited + 1)	(2) log(purchased + 1)	(3) waited pr	(4) purchased pr
$after \times treated$	0.745***	0.189***	1.402***	0.133***
	(0.054)	(0.054)	(0.066)	(0.035)
$after \times overlap$	-0.411	-0.476	-0.260	-0.352
	(0.347)	(0.363)	(0.297)	(0.288)
R-squared Adj.	0.211	0.067	0.214	0.002
N Obs	39870	39870	39870	39870
N Series	633	633	633	633

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 10: Check for potential violation of SUTVA using reader base overlap

5.2.4 Ruling Out Alternative Explanations

Next, we explore whether the estimated treatment effect could arise from the platform systematically coordinating the wait-time reduction with new episode releases or other marketing activities such as sending episode gifts and promotions. We provide evidence supporting that the treatment effect we observe is due to the wait-time reduction and unlikely to be primarily driven by other correlated factors.

One alternative explanation for the positive treatment effect is that the platform sends episode gifts or conducts promotions in anticipation of or simultaneously with wait-time reduction to draw reader attention. If this were the case, our estimates might reflect the effect of those marketing activities rather than the pure effect from shorter wait-times. Although we control for these factors in our main analysis, we re-estimate the model by completely removing treated and non-treated series that had any promotions or gifts within our 30-day timeframe, leaving 172 treated series (out of 191) and their matched series. Results are presented in the top half of Table 11 and the effects hold.

Another explanation is that the timing of wait-time reduction is coordinated with new episode releases, in which case the estimated effect would again be confounded by renewed reader interest in the new episode. Again, although we control for days since last episode publication in our main analysis, we re-estimate the model by completely removing treated and non-treated series that had any new episodes released within our 30-day timeframe, leaving 133 treated series and their matched series. Results are presented in the bottom half of Table 11 and the effects hold.

	(1) log(waited + 1)	(2) log(purchased + 1)	(3) waited pr	(4) purchased pr
$after \times treated$	0.732***	0.220***	1.382***	0.178***
(excl. marketing)	(0.055)	(0.057)	(0.070)	(0.040)
R-squared Adj.	0.095	0.012	0.214	0.002
N Obs	31920	31920	31920	31920
N Series	551	551	551	551
$after \times treated$	0.762***	0.199***	1.487***	0.164***
(excl. new episode)	(0.059)	(0.061)	(0.071)	(0.042)
R-squared Adj.	0.227	0.067	0.244	0.002
N Obs	26700	26700	26700	26700
N Series	503	503	503	503

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 11: Robustness Checks on Marketing Activities and New Episode Releases

5.2.5 Alternative Model Specifications

We also estimate the main model with alternative dependent variable transformations (Box-Cox and Log base 10), time periods (14 and 60 days around treatment) and the number of matched controls on propensity scores (1-10 matches). The results from these models replicate the findings from Table 6 and appear in Web Appendix.

6 Exploring the Mechanism

Having demonstrated that reducing wait-times not only increase waited consumption but also episode purchases, we now turn our attention to exploring why such effect arises. To build intuition, we revisit the unique features of serialized information goods discussed in the introduction. Mittell (2015) defines a series as a "cumulative narrative that builds over time, rather than resetting back to steady-state equilibrium at the end of every episode." Each episode in a series might have its own mini-resolution, but the overarching narrative continues and gradually unfolds throughout the entire series. Naturally, there is a defined sequence of consumption, and the build-up over episodes can be fully appreciated by only those who have consumed the earlier episode (Linkis, 2021), which we referred to as directed complementarity. Moreover, the complementarity value diminishes in time as the audience forget and lose emotional momentum.

As consuming episodes is not costless (e.g., time, effort, money), the consumer's decision on whether and how to consume depends on her net benefit (valuation for the episode minus cost). We also empirically showed in Section 3.2 that users read an episode only if they have read the preceding episode, which implies that in this case the complementarity value is required for the episode to yield positive utility. Conversely, if a consumer decides not to consume an episode, she will not consume any of the episodes later in the sequence, even though she might have purchased some of those episodes had she had the chance. This is a major impediment to sales since the platform loses out on the sales of not only that episode, but also all following episodes. As a concrete example, the audiobook platform Storytel reverted from offering ten-episode serial format to a single publication, citing the problem that the "consumers jump off, they kind of get nine chances to jump off the series" (Berglund and Linkis, 2022).

Reducing wait-times serve as a solution to this problem by allowing consumers to consider the purchase decision for more episodes. Shorter wait-time induces consumers with lower preference to switch from no consumption to waiting because the valuation of the immediate episode is higher *and* it allows her to consider consuming the next episode, which now has become more likely that she will also consume. Under uncertainty of her preference for the next episode, she has higher incentive to at least "check it out." Then, the platform has a chance that she enjoys the immediate episode and cannot help but pay to consume the next episode immediately. Moreover, since consumers expect to consume more episodes ex ante, more of them will choose to *start* the series for the first time. Finally, because the valuation of waited episodes are now higher, the consumers may consume the waited episodes at a faster rate. In sum, shorter wait-times may cause more consumers to consume more episodes and do so at a faster rate, ultimately leading to increased revenues.

We next provide a simple model framework that conceptualizes the episode consumption decisions in a series according to the setting outlined above. We then formalize the potential mechanisms that may explain our results and take them to our data.

6.1 Conceptual Framework

In our model, a monopoly firm sells a series that consists of two episodes, where each episode is offered in two versions: the high quality (no wait-time) q_H offered at price p and the low quality (with wait-time) q_L ($q_L < q_H$) offered for free ($p = 0$). Without loss of generality, we set $q_H = 1$ and $0 < q_L < 1$. Each consumer has a demand for at most one unit of each episode. To capture heterogeneity in consumer taste across episodes, we denote consumer i's base preference for episode $e \in \{1, 2\}$ of series s as θ_{ise}. Assume that consumers know their preference for the first episode, θ_{is1}, but their preference for the second episode, θ_{is2}, is revealed only after they consume the first episode. Instead, the consumers only know the distribution of θ_{ise} for a given series s, denoted as F_s defined on $[0, 1]$. This reflects the idea that a consumer's preference for an episode, which is an experience good, is determined by how much she enjoyed the preceding episode of the series. Without having read the preceding episode, she only knows the distribution based on her beliefs of the overall quality of the series.

We assume consumer utility is a linear function of the valuation, and for juxtaposition, let us first analyze the simplified setting where the firm sells only the first episode (standalone good). Dropping subscripts for consumer i and series s, the consumer obtains utility $U_{p1} = \theta_1 - p - r$ from purchasing and utility $U_{w1} = \theta_1 q_L - r$ from waiting. r represents the non-monetary cost of reading an episode such as time and effort spent, which ensures that some consumers choose to not consume even for free. Versioning with wait-times is feasible because consumers prefer to consume the episode sooner but vary in the degree of preference. Denote $\theta_H = \frac{p}{1-q_L}$ to be the marginal consumer indifferent between purchasing and waiting and $\theta_L = \frac{r}{q_L}$ to be the marginal consumer indifferent between waiting and no consumption. Readers in $[\theta_L, \theta_H)$ will wait to read for free, $[\theta_H, 1]$ will purchase to read immediately, and $[0, \theta_L)$ will not read the episode. In this simplified single-episode case, if the firm increases q_L (i.e., lower wait-times), θ_L will decrease (consumption expansion effect), and θ_H will increase (cannibalization effect). In other words, at the low end, more readers will switch from no consumption to waiting, and at the middle, some consumers will switch from purchasing to waiting. Moreover, because waited episodes yield no proceeds for the firm, aggregate revenues for the episode will unequivocally decrease.

We now turn to the two-episode setting. First, due to the defined sequence and directed complementarity, if one does not consume Episode 1, she churns from the series and does not consume Episode 2. Hence, readers with low θ_1 would not consume either episode, although some might have drawn a high value of θ_2. Second, the consumer's utility from the second episode increases if she has read the first episode due to complementarity, denoted as C. Moreover, the complementarity value decreases as the time interval between successive consumption instances increase. In other words, consumers receive the highest complementarity value with the second episode from purchasing and a lower value from waiting depending on the wait-time.

Hence, in the consumption decision for Episode 1, the reader not only considers her preference for the immediate episode (θ_1), but also the *expected utility* from Episode 2 conditional on θ_2. The consumer obtains utility $U_{p2} = \theta_2 - p - r + C$ from purchasing and utility $U_{w2} = \theta_2 q_L - r + C q_L$ from waiting for the second episode. Hence, the expected utility from the second episode is as follows:

$$\begin{aligned}\mathrm{E}(U_2) = \mathrm{E}(U_{w2} \mid \frac{r}{q_L} - C < \theta_2 < \frac{p}{1-q_L} - C) \cdot Prob(\frac{r}{q_L} - C < \theta_2 < \frac{p}{1-q_L} - C) \\ + \mathrm{E}(U_{p2} \mid \frac{p}{1-q_L} - C < \theta_2) \cdot Prob(\frac{p}{1-q_L} - C < \theta_2) \quad (7)\end{aligned}$$

For Episode 1, the consumer chooses the option that will maximize her total expected utility: $U_{p1} = \theta_1 - p - r + \mathrm{E}(U_2)$ from purchasing; $U_{w1} = \theta_1 q_L - r + \mathrm{E}(U_2)$ from waiting; and zero utility from no consumption. The marginal consumer $\theta_H = \frac{p}{1-q_L}$ is the same as in the single-episode setting, but $\theta_L = \frac{r - \mathrm{E}(U_2)}{q_L}$ is now lower. In words, consumers with lower preference for the first episode choose to consume nonetheless since it enables them to consume the second episode.

What are the implications on the consumption decision when the firm increases q_L? θ_L decreases under shorter wait-times, which allows a greater mass of the reader base with lower preference (θ_1) to *consider* consuming the second episode. Moreover, the decrease in θ_L would be greater when the distribution of θ, F_s, is concentrated around the middle as the increase in U_{w2} from Equation 7 would be applicable to a larger proportion of the distribution. This implies that the consumption expansion effect is greater when the overall series is considered moderately good. The intuition is as follows: when a consumer expects episodes in the series to be of high quality (right skewed F_s), due to for example suspenseful cliff-hangers or a captivating plot, she expects to draw a high value of θ_2 and purchase the second episode. The increase in q_L (wait-time reduction) in turn has limited upside on the expected utility from the second episode, and thus leads to less consumers switching from no consumption to waiting. Similarly, when a consumer expects the series to be not enjoyable (left skewed F_s), she expects to draw a low value of θ_2 and likely not consume. Again, the reduced wait-time has little effect on the consumption decision of the first episode.

Based on the above framework, we formalize three mechanisms that may drive our main results. First, shorter wait-times induce a reader to progress further in the series, and in doing so, she ends up purchasing more episodes, which may offset cannibalization. Intuitively, a reader who has low preference for the immediate episode may switch from no consumption to waiting because she has to wait less *and* she expects to receive higher utility from the following episode. If this were true, the positive change in progress would be more pronounced for series that are considered moderately good. Having said that, what happens to total purchases made by a reader is an empirical question. Readers now face more opportunities to make purchases, but they are also less incentivized to purchase.

Second, shorter wait-times cause more readers to start the series for the first time. A consumer decides to start a series if the aggregate expected utility from the episodes outweighs the cost associated with starting a new series. The total expected utility from the series increases as the reader expects to progress further under reduced wait-times, and more consumers will now choose to start the series, serving as additional sources of revenue. As the increase is more pronounced for series with likelihood of waiting, we expect an inverted-U shaped relationship between the inflow of new readers and series quality.

Finally, shorter wait-times increase the pace at which readers consume the waited episodes. When a consumer chooses to wait, it is in the platform's best interest for her to consume the episode as soon as the wait-time elapses so that she can potentially purchase the next. Under shorter wait-times, waited consumption yields higher utility, both from the immediate episode and the subsequent episode, providing greater incentives for the readers to consume as soon as the wait-time elapses. For example, a reader is more excited to read an episode after having waited for an hour versus a day, and by consuming this episode, she only has to wait for an hour to consume another free episode. Similar to the above two mechanisms, the increase in expected utility would be the greatest if the reader expects to wait for the subsequent episode. Hence, we again expect an inverted-U shaped relationship between the increase in the pace of claiming waited episodes and series quality.

6.2 Empirical Validation of Proposed Mechanisms

6.2.1 Further Progress within the Series

We empirically examine whether readers facing shorter wait-times indeed consume more episodes in a series before churning. Shorter wait-times not only make the immediate episode more attractive, but also increase the likelihood of consumption and thereby the expected utility from later episodes, as shown in Equation (7). Series believed to be moderately good would be more heavily impacted as consumers ex ante expect to wait for the subsequent episodes. Then, we investigate whether readers also purchase more episodes in the series. This is an empirical question, since there are two countervailing effects: readers have more opportunities to make purchases, but are also more incentivized to wait rather than purchase.

To obtain groups of similar users who face different wait times for the same series, we begin by sampling two groups of readers for each treated series and its matched control series. The groups includes users who first started reading the series within 30 days before (and after) treatment, which we call the *before group* (and the *after group*). The assumption is that given the tight 30-day time frame, any differences in individual consumption patterns for the treated series, relative to the matched control series post treatment, can be attributed to the wait-time reduction. We also confirm the robustness of this effect to changes in the window to 15 days. The dataset was filtered for series that have at least 10 new readers in each group to ensure sufficient sample size, leaving 137 treated series and their matched control series. Then, we measure the total number of consumed and purchased episodes of each series for each reader. For a clean comparison of behaviors under different wait-time regimes, we drop readers that are observed to continue reading the series either after the treatment date (*before group*) or 30 days after treatment date (*after group*). Hence, we have a clean comparison of readers that started and churned from a series just before and after the wait-time reduction. For robustness, we check whether including these readers makes a difference and find no qualitative difference.

Since we cannot precisely quantify series quality (i.e., the distribution of the consumer's belief for episodes in the series, F_s), we leverage two measures that serve as proxies: the number of average daily unique readers (*Avg. DUR*) and the proportion of waited episodes relative to total consumed (*Prop. Waited*). Presumably, a large reader base would indicate that a high proportion of readers have high preference for the episodes (right skewed F_s), whereas a small reader base would indicate the opposite (left skewed F_s). We compute *Avg. DUR* by taking the average of daily unique reader count over the 15 days prior to the wait-time reduction. According to our framework, we expect an inverted-U shaped relationship between episodes consumed and *Avg. DUR*. Alternatively, a higher *Prop. Waited* would indicate a higher distribution of F_s concentrated in the middle, causing more consumers to wait than purchase. In this case, we expect the impact on episodes consumed to increase with *Prop. Waited*.

We measure the average effect of wait-time reduction on individual consumption by estimating Equation 4 with two periods, adding interaction terms between the treatment indicator and *Avg. DUR* and its squared term (or *Prop. Waited*). The estimation results in Table 13 yield results consistent with the above mechanism. Column (2) demonstrates an inverted U-shaped relationship between the effect size and the average daily unique readers, and column (3) shows a positive, significant coefficient for proportion of waited consumption. Columns (4)-(6) indicate that there is no significant change in total episodes purchased, as the positive consumption expansion effect across episodes offset the negative cannibalization effect. In words, although a reader is less incentivized to purchase a given episode under reduced wait-times, she faces more opportunities to purchase as she progresses further in the series, ultimately making a similar number of purchases.

	log(Consumed + 1)			log(Purchased + 1)		
	(1)	(2)	(3)	(4)	(5)	(6)
$after$	0.051	0.072	0.052	0.050	0.031	0.046
	(0.053)	(0.049)	(0.052)	(0.043)	(0.040)	(0.042)
$after \times treated$	0.472***			-0.119		
	(0.087)			(0.077)		
$after \times treated \times Avg.\ DUR$		1.25e-03***			1.44e-07	
		(0.000)			(0.000)	
$after \times treated \times (Avg.\ DUR)^2$		-3.34e-07***			3.10e-08	
		(0.000)			(0.000)	
$after \times treated \times Prop.\ Waited$			0.760***			-0.131
			(0.137)			(0.130)
R-squared Adj.	0.005	0.005	0.005	0.000	0.000	0.000
N Obs	256341	256341	256341	256341	256341	256341
N Series	398	398	398	398	398	398
Series group-series FE	Y	Y	Y	Y	Y	Y

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 12: Treatment Effect on Individual Consumption Moderated by Proxies for Series Quality

6.2.2 Increased Inflow of New Readers

We found that the reader is more likely to progress further in a series under shorter wait-times, which would increase the a priori value of the series. In turn, we expect more readers to start the series, expanding the

	log(Consumed + 1)			log(Purchased + 1)		
	(1)	(2)	(3)	(4)	(5)	(6)
$after$	0.022	0.044	0.024	0.063	0.043	0.058
	(0.054)	(0.051)	(0.054)	(0.042)	(0.039)	(0.041)
$after \times treated$	0.491***			-0.108		
	(0.082)			(0.077)		
$after \times treated \times Avg.\ DUR$		1.26e-03***			4.73e-05	
		(0.000)			(0.000)	
$after \times treated \times (Avg.\ DUR)^2$		-3.42e-07***			1.45e-08	
		(0.000)			(0.000)	
$after \times treated \times Prop.\ Waited$			0.785***			-0.114
			(0.127)			(0.127)
R-squared Adj.	0.010	0.009	0.010	0.007	0.007	0.007
N Obs	266985	266985	266985	266985	266985	266985
N Books	398	398	398	398	398	398
Group-Book FE	Y	Y	Y	Y	Y	Y

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 13: Treatment Effect on Individual Consumption Moderated by Proxies for Series Quality

total size of the pie. Analogous to retailers incentivizing store visits, the series lures in a larger traffic of new readers in the hopes that they will purchase episodes conditional on starting the series. Moreover, because the incremental progress is greater for series of moderate quality, we expect to see a similar moderating effect of proxies for series quality on the inflow of new readers. To see the empirical impact of new readers, we estimate Equation 4 using the daily count of new readers as the dependent variable and the treatment term interacted with proxies for series quality. Results displayed in Table 14 are consistent with the mechanism. Column (1) demonstrates that the wait-time reduction leads to a positive, significant change in the inflow of new readers. The moderator coefficients estimated in columns (2) and (3) provide evidence that the effect has an inverted U-shaped relationship with *Avg. DUR* and is increasing with *Prop. Waited*.

6.2.3 Reduced Excess Wait-time

Platforms focus on maintaining regular and frequent user engagement with the product as it is a critical driver of customer lifetime value. The WFF policy incentivizes regular consumption by setting wait-times conditional on accessing the last waited episode. Recall that the wait-time clock starts to tick from the moment that the reader claims her free episode. This unique aspect motivates the reader to visit the series once she is eligible for a free waited episode. In order to establish frequent engagement, the platform must ensure that the reader accesses her free episode shortly after she becomes eligible.

By reducing wait-times, the platform endogenously strengthens the reader's incentive to access the free episode in two ways: increased quality of the immediate episode (e.g., higher excitement, lower forgetfulness), and increased expected utility from the subsequent episode (i.e., access the next episode sooner if she chooses

	log(New Readers + 1)		
	(1)	(2)	(3)
$after \times treated$	0.243***		
	(0.036)		
$after \times treated \times Avg.\ DUR$		9.89e-04***	
		(0.000)	
$after \times treated \times (Avg.\ DUR)^2$		-3.50e-07***	
		(0.000)	
$after \times treated \times Prop.\ Waited$			0.360***
			(0.053)
R-squared Adj.	0.244	0.240	0.243
N Obs	40080	40080	40080
N Series	636	636	636
DOW FE	Y	Y	Y
Series group-series FE	Y	Y	Y
Series group-period FE	Y	Y	Y

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 14: Treatment Effect on Inflow of New Readers Moderated by Proxies for Series Quality

to wait). Hence, we expect readers to be quicker to access the waited episode once the wait-time elapses. Again, we expect to see the highest impact for medium quality series, as the probability of waiting for the subsequent episode is the highest. This acceleration (time-shifting) mechanism leads the readers to make consumption decisions at a faster pace, which can drive the increase in daily aggregate demand.

To empirically validate this mechanism, we measure the change in *excess wait-time* due to wait-time reduction. We define excess wait-time to be the time between when an episode becomes eligible to be consumed for free for a given reader and when she actually consumes the episode. If our mechanism holds, we would expect the excess wait-time to decrease after treatment. We begin by sampling waited consumption 30 days before and after treatment date for each treated series and its matched control series. We then compute the excess wait-time based on the time of access for the preceding episode of the same series.

We measure the average effect of wait-time reduction on excess wait-time (log-transformed) by estimating Equation 4 with two periods, adding interaction terms between the treatment indicator and $Avg.\ DUR$ and its squared term (or *Prop. Waited*). The estimation results in Table 15 are lend support for the mechanism. As shown in column (1), there is a significant decrease in excess wait-time under shorter wait-times. Column (2) present an inverted-U shaped relationship between the effect size and $Avg.\ DUR$, and column (3) shows a linearly increasing relationship with *Prop. Waited*.

6.3 Heterogeneous Treatment Effects

There are sizeable effects of reducing wait-time on individual progress within the series, inflow of new readers and excess wait-time. We now explore how these effects differ across consumers with purchase propensity as

	log(Excess Wait-time + 1)		
	(1)	(2)	(3)
$after$	0.049	0.041	0.048
	(0.037)	(0.036)	(0.037)
$after \times treated$	-0.240***		
	(0.076)		
$after \times treated \times Avg.\ DUR$		-6.35e-04***	
		(0.000)	
$after \times treated \times (Avg.\ DUR)^2$		1.95e-07***	
		(0.000)	
$after \times treated \times Prop.\ Waited$			-0.330***
			(0.112)
R-squared Adj.	0.001	0.001	0.001
N Obs	14619682	14619682	14619682
N Series	626	626	626
Series group-series FE	Y	Y	Y

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 15: Treatment Effect on Excess Wait-time Moderated by Proxies for Series Quality

the moderator. Take an extreme example of a reader that chooses only between waiting and no consumption and another that chooses only between purchasing and no consumption. Intuitively, reduced wait-times should only impact the former while having no impact on the latter. Formally, we extend the conceptual framework from Section 6.1 by introducing individual heterogeneity along the price coefficient, denoted by $\alpha_i \in \{\alpha_L, \alpha_H\}$ ($\alpha_L < \alpha_H$). α_i is an inherent characteristic fixed for consumer i, and *H-type* is more likely to wait than purchase compared to *L-type*. The consumer's utilities from purchasing are modified as $U_{p1} = \theta_1 - \alpha p - r$ and $U_{p2} = \theta_2 - \alpha p - r + C$, with the only change being the addition of α as the price coefficient. The utilities from waiting the same as before. Then, following the same argument, it is straightforward to see that increasing q_L leads to a greater increase in $\mathrm{E}(U_2)$ and hence a greater consumption expansion effect for *H-type* consumers. Note that α_i is a factor that dictates one's time-money trade-off and may represent both price sensitivity and the level of patience. We cannot separately identify the two as it would affect consumption behavior in the same way.

We estimate the heterogeneous treatment effects by using a median split on readers' purchase propensity. We label every reader in the dataset as either *L-type* or *H-type* based on purchase propensity, computed as the ratio of episodes purchased to total consumed (waited and purchased) over the entire 14-month observation period across all series. Readers are labeled as *H-type* if purchase propensity is below the median (0.337) and *L-type* if above.

Table 17 presents the analysis from Section 6.2.1 using purchase propensity as the moderator. As expected, readers that tend to purchase more progress further in the series and make more purchases (coefficients on *L-type*). The heterogeneous treatment effect on total episodes consumed shown in column (2) is consistent with our expectations: the effect of wait-time reduction is 0.536 for *H-type* readers, but 21% lower (-0.112) for *L-type* readers. In words, the incremental progress is markedly higher for readers that tend to wait. Having said that, column (4) reports a greater effect on episodes purchased for *L-type*. Although *H-type* readers consume much more under shorter wait-times, there is no significant change in purchases. Put differently, cannibalization is offset by purchases made on incremental episodes consumed. On the other hand, the *L-type* progress further to a lesser extent, but because they are more inclined to make purchases, the number of purchases significantly increase in aggregate.

	log(Consumed + 1)		log(Purchased + 1)	
	(1)	(2)	(3)	(4)
L-type		0.208***		1.069***
		(0.034)		(0.053)
$after$	0.051	0.050	0.050	0.047
	(0.053)	(0.053)	(0.043)	(0.038)
$after \times treated$	0.472***	0.536***	-0.119	-0.106
	(0.087)	(0.088)	(0.077)	(0.067)
$after \times treated \times$ *L-type*		-0.112***		0.221***
		(0.043)		(0.067)
R-squared Adj.	0.005	0.012	0.000	0.170
N Obs	256341	256341	256341	256341
N Series	398	398	398	398
Series group-series FE	Y	Y	Y	Y

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 16: Heterogeneous Treatment Effect on Individual Consumption Moderated by Purchase Propensity

	log(Consumed + 1)		log(Purchased + 1)	
	(1)	(2)	(3)	(4)
H-type		0.203***		1.028***
		(0.034)		(0.052)
$after$	0.022	0.020	0.063	0.050
	(0.054)	(0.054)	(0.042)	(0.037)
$after \times treated$	0.491***	0.552***	-0.108	-0.102
	(0.082)	(0.082)	(0.077)	(0.067)
$after \times treated \times$ *H-type*		-0.104**		0.239***
		(0.043)		(0.073)
R-squared Adj.	0.010	0.017	0.007	0.166
N Obs	266985	266985	266985	266985
N Books	398	398	398	398
Group-Book FE	Y	Y	Y	Y

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 17: Heterogeneous Treatment Effect on Individual Consumption Moderated by Purchase Propensity

Table 18 presents the analysis from Section 6.2.2 separately showing the effect on new reader inflow for *L-type* and *H-type*. While the effect of wait-time reduction on the daily inflow of new readers is positive and significant for both types, we see that the effect is much larger for *H-type* readers (0.241 vs. 0.101). The increase in expected utility from the series is higher for consumers that tend to wait as they will face shorter wait-times and thus expect to benefit more, again providing empirical support for our framework.

	log(New Readers + 1)		
	All Readers	*L-type*	*H-type*
$after \times treated$	0.243***	0.101***	0.241***
	(0.036)	(0.027)	(0.033)
R-squared Adj.	0.244	0.204	0.263
N Obs	40080	40080	40080
N Series	636	636	636

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 18: Treatment Effect on New Readers by Type

Finally, Table 19 presents the analysis from Section 6.2.3 using purchase propensity as the moderator for the effect on excess wait-time. Again, the heterogeneous treatment effects shown in column (2) are aligned with our expectations. The coefficient on *L-type* is negative and significant (-0.169), meaning those readers on average unlock their waited episodes sooner than the patient readers. The effect of wait-time reduction is -0.253 for *H-type* readers, but 49% lower (0.124) for *L-type* readers. It is more likely that a reader who tends to wait will also choose to wait for the subsequent episode, and hence, reducing the wait-time further incentivizes her to unlock the immediate waited episode as soon as she is able to do so.

	log(Excess Wait-time + 1)	
	(1)	(2)
L-type		-0.169***
		(0.034)
$after$	0.049	0.049
	(0.037)	(0.037)
$after \times treated$	-0.240***	-0.253***
	(0.076)	(0.076)
$after \times treated \times$ *L-type*		0.124***
		(0.047)
R-squared Adj.	0.001	0.001
N Obs	14619682	14619682
N Series	626	626
Series group-series FE	Y	Y

Note: Robust standard errors clustered at series level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 19: Heterogeneous Treatment Effect on Excess Wait-time Moderated by Purchase Propensity

7 Conclusion

Digital content platforms today are not only competing among peers of the same medium but also across medium – content produced in text, video and audio forms are all fighting to occupy consumers' free time such that they will be the first in line when consumers open up their smartphones. In such a competitive environment, serialization has become the prevalent mode of publication due to its effectiveness in encouraging consistent and recurring consumption. Moreover, the ability of digital platforms to track individual consumption has allowed them to exploit time as a monetization lever, leading to the adoption of "wait-for-free" (WFF). In this research, we ask whether giving more for free (i.e., reducing wait-times) can increase platform revenues in the context of serialized information goods. In the absence of factors such as advertising revenues, network effects or sampling that justify inducing more free consumption, it appears a unilateral reduction in wait-times would only lead to revenue cannibalization.

Using data from a platform serving serialized books, we leverage an exogenous policy change where the platform reduced the wait-times for a set of series. We estimate using a difference-in-difference framework how the wait-time reduction impacted the number of waited or purchased episodes. We find that, on average, the demand for free and paid consumption increased following the reduction. We conduct a battery of robustness checks to rule out any spurious correlations and validate the causal relationship.

We then explore the mechanisms at work. We begin with a conceptual framework to analyze consumption decisions in a simple two-episode series. Importantly, the framework incorporates three unique features of serialized information goods: defined sequence of consumption, complementarity between episodes and diminishing complementarity value over time. These unique features shed light on three potential mechanisms that may be driving our results, for which we provide empirical evidence from the data. First, the platform retains consumers over a longer stretch of episodes, allowing the platform to monetize each consumer to a greater extent in aggregate. Second, the a priori value of the series increases for the consumers, thereby leading to an increase in new consumers. Third, the value of waited episodes increases under shorter wait-times, which accelerates the rate of consumption. In short, the platform benefits from retaining a larger consumer base over a longer period that may generate revenues at a faster pace.

Our results offer managerial insights relevant for serialized content platforms. First, they confirm that a counter-intuitive approach of allowing more free consumption by means of reduced wait-time can indeed increase demand for paid episodes. Second, we document that the approach not only attracts new consumers, but also can monetize existing consumers to a greater extent (i.e., cannibalization is transient). Third, our results suggest that series considered to be of moderate quality would benefit the most from the wait-time reduction. Fourth, our findings indicate heterogeneous effects across consumers based on purchase propensity. For the segment of price sensitive or patient consumers, it may not be worthwhile to allow easier consumption, which suggests that the platform may benefit from a targeted policy.

Although the present study is one of the first to investigate the novel WFF policy and the economics of serialized information goods, it is not without its limits. First, our research relies on non-experimental variation in the data. Although we leverage various identification strategies to support causality, a randomized field experiment that manipulates wait-times across consumers for a given series could further strengthen our findings. Second, our results are specific to our data context. Future research explore whether they also hold for other types of serialized information goods such as videos and whether the results indeed point in the opposite direction when wait-times are increased. Third, our analysis demonstrates the causal effect of changing wait-times, but cannot comment on the optimal wait-time, which would require estimating a structural model of consumers' episode consumption. This would also allow one to investigate counterfactual policies such as charging a positive price on waited episodes or targeted wait-times. Estimating a state-dependent utility model as a function of individual fixed effect, consumption context and episode content to explore optimal policies would be a fruitful research area. Leveraging recent advancements in text analysis, it may be worth inquiring how the episode content such as the strength of cliffhangers, level or suspense and sentiment affects consumption decisions. Fourth, although we provide ample evidence that correlated marketing activities are unlikely the driver of our results, we cannot fully rule out that in some instances firms may still have tried to support the treated series with alternative methods such as curation or recommendation algorithms, which might have had some effect on our estimates. Fifth, we focus on the short-term effect of the wait-time reduction, as identification of a long-term effect is more difficult. It would be interesting to explore the long-term effect of varying wait-times on consumers' long-term behavior on the entire platform and subsequent series they later consume.

8 Appendix

8.1 Overview of Robustness Checks

Concern		Proposed Robustness Check
1)	Treatment effect may be coincidentally picking up some spurious effects	Falsification tests using (1) pseudo treatment indicators by randomly assigning a control series as treated; and (2) pseudo treatment dates by manipulating the treatment date to be 15 days prior to the actual treatment date.
2)	Wait-time reduction on one book may affect demand for adjacent books, violating SUTVA	Test for spillover effects by estimating the model including percentage overlap of readers between treated and control series, interacted with the post-treatment indicator. To be conservative, we allow up to 15 matched control series per treated series.
3)	Despite controlling for observable factors via PSM, treated and untreated books may differ on unobserved factors	Heckman two-stage selection model, using *program type* of the series to satisfy exclusion restriction.
4)	Results may be biased if the platform strategically reduced wait-times around the time of episode updates or marketing activities	Estimate the model excluding series which had updates, gifts or promotions in the 30 days around the wait-time reduction.
5)	Results may not be robust to the functional forms	Estimate the model using Box-Cox and $\log_{10}$ transformations of the dependent variable.
6)	Results may not be robust to different time frame or number of PSM matches	Estimate the model with different timeframes (7, 15, 30 days before/after treatment) and number of matches (1-10).

Table 20: Overview of Robustness Checks

References

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

Ascarza, E., Netzer, O., and Runge, J. (2020). The twofold effect of customer retention in freemium settings. *Columbia Business School Research Paper Forthcoming.*

August, T., Dao, D., and Shin, H. (2015). Optimal timing of sequential distribution: The impact of congestion externalities and day-and-date strategies. *Marketing Science*, 34(5):755–774.

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

Berglund, K. and Linkis, S. (2022). Modelling subscription-based streaming services for books. *Mémoires du livre*, 13(1):1–30.

Borusyak, K. and Jaravel, X. (2018). *Revisiting event study designs*. SSRN.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72.

Calzada, J. and Valletti, T. M. (2012). Intertemporal movie distribution: Versioning when customers can buy both versions. *Marketing Science*, 31(4):649–667.

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

Chiou, L. and Tucker, C. (2013). Paywalls and the demand for news. *Information Economics and Policy*, 25(2):61–69.

Choi, A. A., Rhee, K.-E., Yoon, C., and Oh, W. (2023). The cost of free: The effects of "wait-for-free" pricing schemes on the monetization of serialized digital content. *MIS Quarterly*, 47(3).

Choi, J., Chae, I., and Feinberg, F. M. (2022). Wait for free: A consumption-decelerating promotion for serialized digital media. *Available at SSRN: https://ssrn.com/abstract=4098036*.

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

Datta, H., Knox, G., and Bronnenberg, B. J. (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 D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96.

Deneckere, R. J. and McAfee, P. R. (1996). Damaged goods. *Journal of Economics & Management Strategy*, 5(2):149–174.

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

Deshpande, M. and Li, Y. (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 Sekhon, J. S. (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.

Eco, U. (1990). *The Limits of Interpretation*. Indiana University Press.

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

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

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

Gu, X., Kannan, P., and Ma, L. (2018). Selling the premium in freemium. *Journal of Marketing*, 82(6):10–27.

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.

Haviland, A., Nagin, D. S., and Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological methods*, 12(3):247.

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

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161.

Imai, K., Kim, I. S., and Wang, E. H. (2021). Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*.

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

Kamada, Y. and Öry, A. (2020). Contracting with word-of-mouth management. *Management Science*, 66(11):5094–5107.

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

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

Lee, C., Kumar, V., and Gupta, S. (2019). Designing freemium: Strategic balancing growth and monetization. *Available at Github: https://vineetkumars.github.io/Papers/DesigningFreemiumS2019.pdf.*

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

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

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

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

Mittell, J. (2015). *Complex TV: The poetics of contemporary television storytelling*. nYU Press.

Mussa, M. and Rosen, S. (1978). Monopoly and product quality. *Journal of Economic theory*, 18(2):301–317.

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

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

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

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

Shi, Z., Zhang, K., and Srinivasan, K. (2019). Freemium as an optimal strategy for market dominant firms. *Marketing Science*, 38(1):150–169.

Stigler, G. J. and Becker, G. S. (1977). De gustibus non est disputandum. *The american economic review*, 67(2):76–90.

Stuart, E. A., Lee, B. K., and Leacy, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology*, 66(8):S84–S90.

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

Yoganarasimhan, H., Barzegary, E., and Pani, A. (2022). Design and evaluation of optimal free trials. *Management Science*.

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

Zhang, Z., Kim, H. J., Lonjon, G., Zhu, Y., et al. (2019). Balance diagnostics after propensity score matching. *Annals of translational medicine*, 7(1).

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