Monetizing Serialized Content:

How "Wait for Free" Impacts Paid and Free Consumption

Peter S. Lee, Vineet Kumar, K. Sudhir

April 30, 2024

Abstract

Entertainment such as novels and audio/video content are increasingly published as serialized chapters/episodes on digital media platforms. As a monetization strategy, these platforms offer a "wait-for-free" (WFF) option where consumers can consume the next episode in a sequence immediately after paying a per-episode cost or for free after waiting for a specified wait-period. Despite the growing interest in leveraging time as a discrimination device for serialized media, there is limited research on how time affects consumption decisions of serialized content. Hence, we pose the following simple but critical question: Can allowing quicker free consumption increase revenues? While it is easy to see that a shorter wait to access free material will attract new consumers, a natural concern is that it will also cannibalize paid consumption from existing consumers. In this paper, we argue that unique characteristics of serialized content – complementarity in utility from consuming sequential episodes, the diminishing value of complementarity over time and the consumer's time-varying willingness-to-pay – can actually lead to an increase in paid consumption among existing customers. By leveraging a natural experiment from a serial fiction platform, we estimate the impact of reducing wait-times on downstream consumption behaviors using a difference-in-difference framework, while addressing potential selection issues using a matching-based approach. We find that existing consumers increase paid consumption by 12%. Together with incremental purchases from new consumers who start consuming the series after the reduction, aggregate paid consumption increases by 19%.

Keywords — serialized media, wait-for-free (WFF), product versioning

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 media refer to information goods 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 the way for serialization as a popular mode of publication across product types. Partitioning a unified content into short episodes has become common as it fits 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). The largest platform for serialized fiction novels, Wattpad, is reported to now have over 80 million readers; the leading serialized comics platform, Webtoon, boasts over 85 million users. The video streaming giants Netflix, Amazon Prime Video and Disney+ together serve more than 600 million subscribers.

Many platforms with serialized content across a variety of product domains offer consumers a "wait-for-free" (WFF) option, such as Webtoon (comics), Radish Fiction (books), ReelShort (videos) and Real Racing (games). With WFF, consumers make consumption decisions for each episode: they can access an episode immediately for a fixed price or for free after waiting for a pre-specified wait-time since accessing the prior episode of the same series. Essentially, the WFF option is a form of versioning where consumer can choose between two vertically differentiated episodes—one with zero waiting times for a price and another with a non-zero waiting time for free. Consumers who have high willingness-to-pay (WTP) for an episode may pay for immediate consumption, while those with lower WTP can choose to have free but delayed consumption.[1]

The cross-sectional impact of versioning across different types of consumers on profits/revenues through the benefits of new user acquisition and the cost of cannibalization has received significant attention in the literature. For example, the literature on freemium products discuss a variety of mechanisms through which the relative positioning of the products affect consumption and revenues. Examples include product adoption through referrals (Lee et al., 2019; Kamada and Öry, 2020), network externalities (Shi et al., 2019), advertising revenues (Chiou and Tucker, 2013; Lambrecht and Misra, 2017) and consumer learning (Li et al.,

[1] At first glance, the time based versioning with serialized media has similarities with choices among higher and lower priced versions of products that are sequentially released (e.g., hard versus soft cover books, movies at theaters versus streaming/DVDs as modeled in Luan and Sudhir 2022). We characterize the conceptual differences arising from partitioning of the unified content into serialized content and their implications for consumer behavior and monetization as we motivate our research questions.

2019; Deng et al., 2022). Increasing the value of the free option dissuades existing consumers from purchasing, but at the same time attracts new consumers who may choose to purchase.

With WFF, the use of time as a discrimination device opens up new possibilities of *intertemporal discrimination within consumers*. If the willingness-to-pay for the same product varies across time for a given consumer, the firm can segment a consumer into states with high and low WTP to extract higher revenues. By imposing wait-times relative to the timing of individual consumption (unlike static release schedules studied in Luan and Sudhir 2022 and Zhao et al. 2022), the WFF policy repeatedly offers two vertically differentiated options such that the consumer self-selects into the option based on her time-varying WTP.

Although the WFF option for serialized media is now widely used, there is limited research that sheds light on how time affects consumer decisions, with a focus on the role of serialization. A notable exception is Choi et al. (2022); the paper studies how WFF impacted free and paid viewing on an online platform for serialized comics. Using a reduced-form model that explains the number of free and paid views aggregated at the episode level, it finds that the introduction of WFF led to an increase in paid episodes views. The paper suggests that WFF generates incremental demand for paid episode views from the new users to the series, which exceeds the effect of purchase cannibalization from the existing users through free viewing.

We build on this nascent research area by examining the downstream impact of wait-times on individual consumption and purchase decisions. We begin by drawing on the unique features of serialized products to conceptualize the consumer's decision making process across time and across episodes. Based on the conceptual framework, we generate hypotheses on various consumption dynamics that arise from changing wait-times, including the depth and pace of consumption and the decision to start a series. Importantly, we argue that owing to the characteristics unique to serialized media, allowing quicker free consumption can actually lead to an increase in paid consumption *even among existing consumers* of the series. We empirically validate our conceptual framework and assess the ensuing impact on aggregate consumption and purchases.

In order to construct our hypotheses, we formalize three foundational characteristics unique to serialized media that affect consumption decisions by synthesizing research across literary studies, marketing and economics. First, episodes exhibit directed complementarities: consuming one episode enhances the value of the next. 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, cutting-off techniques and analepses, allows the audience to slowly build their understanding of and become hooked to the storyline that unfolds (Mittell, 2006). Hence, the consumption of each episode creates a stronger incentive to proceed sequentially in the series (Linkis, 2021).

Second, the value of these complementarities diminishes over the time since the last consumption, aligning with the notion that the subsequent consumption utility decreases as consumption capital dissipates over time (Becker and Murphy, 1988; Heather and Vuchinich, 2003). The internal momentum towards the series gradually decays and consumers forget about the previous episode, which may affect their valuation of the next episode. This time-sensitive aspect of valuation underscores the importance of timing in release strategies for serialized media (Zhao et al., 2022; Godinho de Matos et al., 2023) and the phenomenon of binge consumption (Schweidel and Moe, 2016; Lu et al., 2019, 2023; Godinho de Matos and Ferreira, 2020), where the desire to maintain continuity leads consumers to consume episodes in close succession.

Third, the consumer's willingness-to-pay for an episode exogenously varies over time based on the value of her outside option. The marketing literature on time-based discrimination (sequential release) assume a static WTP within an individual (August et al., 2015; Luan and Sudhir, 2022). Hence, the timing of product release focused on discrimination across consumers with varying WTP. However, because consumers must spend time to consume media products, the urgency to consume depends on the consumption context (e.g., while taking a break vs. working), which is exogenous to the product. This induces the same consumer to make different consumption decisions at different times depending on her state, allowing for discrimination within a consumer across time.

These properties of serialized media create richer consumption dynamics, which enable increased monetization under shorter wait times even among existing consumers of the series. Under shorter wait-times, the consumer gains higher complementarity from waited consumption, which may cause her to switch from purchasing to waiting (cannibalization effect). However, the complementarity properties give rise to positive across-episode spillovers within a consumer that counteract cannibalization through two channels. First, the consumer may switch from no consumption (outside option) or waiting to waiting or purchasing, because both the value of waited consumption for the current episode and the expected value of consumption for the future episodes are higher. Second, this in turn allows her to benefit from complementarity on subsequent episodes, inducing her to make purchases at episodes where her WTP is sufficiently high. In essence, the firm is able to "harvest the acquisition" by retaining consumers over a broader product set.

To comprehensively assess such multifaceted dynamics, we separate out and evaluate the impact on the intensive and extensive margins. As concepts commonly used in the economics literature, intensive margin refers to the magnitude of activity with existing entities (e.g., trade volume, hours of labor), while extensive margin refers to the breadth of participation (e.g., new trade relationships, individuals entering the labor market). We adapt these terms to our specific context: the intensive margin measures how much existing consumers of a series consume or purchase its episodes, and the extensive margin measures how many new consumers begin the series. This distinction allows us to understand how changes in the WFF policy differentially affect the existing consumer base versus attracting new consumers, thereby highlighting the relative contributions of each dimension to aggregate revenue growth.

We empirically assess the consumption impact of WFF using data from a major U.S.-based serialized fiction novel platform that uses the WFF policy. Specifically, we exploit a natural experiment during the period of our data sample, where the platform changed wait-times for a subset of series to help identify the causal effects of changes in wait-times on free and paid consumption.

We use a consumption panel data of over a million users and 20,000 series that covers 15 months from October 2020 to December 2021, which details how the user consumed an episode (i.e., waited or purchased). We augment this dataset with series and episode metadata such as publication date, required wait-time and promotional activities. We also use an in-app currency purchase panel data to explore consumer heterogeneity based on historical platform spending. During the observation period, the platform selected 191 series whose wait-times ranged from 3 to 72 hours and reduced them to a range of 1 to 24 hours. The wait-times pre- and post-reduction as well as the magnitude of reduction vary across the selected series. The change was implemented without prior announcement and aimed to increase user engagement by making it easier for readers to consume episodes for free by waiting. We estimate the average treatment effect of wait-time reduction via a difference-in-differences (DiD) approach, focusing on a 30-day window around the reduction.

Two empirical challenges remain. First, we have to address the potential for selection bias as the series for which wait-times were reduced were decided by the platform. Second, the wait-time reduction was implemented in a staggered manner; hence a standard two-way fixed effects model can produce biased estimates. We address selection bias using a panel matching approach (Imai et al., 2021). We compute the likelihood of a series receiving treatment (i.e., wait-time reduction) at a given time based on observed pre-treatment characteristics. By constructing a control group of series whose wait-times remain unchanged but have comparable probability of being selected for reduced wait-times, we allow the treatment to be random conditional on observable characteristics, satisfying conditional ignorability. We then mitigate potential concerns around staggered adoption by using a stacked DiD approach. The stacked DiD creates event-specific datasets for each of the treated series and its matched control series. By stacking the datasets based on relative periods around the reduction timing and saturating the model with cohort specific fixed effects, we circumvent the issues of biased estimates under staggered treatment adoption (Goodman-Bacon, 2021).

Our empirical analysis proceeds in the following sequence. First, we examine the impact of changing wait-times on the depth of consumption (intensive margin) by examining the change in total number of episodes consumed (i.e., waited and purchased) and purchased by the existing consumers of a series. The estimation results show a 37% increase in episodes consumed per consumer. The key estimate of interest is the impact of wait-time reduction on purchases, as the increased consumption would only hurt platform revenues if it came at the expense of lower purchases. Surprisingly, we find a 12% increase in episodes purchased per consumer, indicating that the positive effect from across-episode spillovers dominates the negative cannibalization effect. Moreover, allowing for heterogeneous treatment effects based on historical platform spending reveals an even greater 17% increase for those that are amenable to paying for content.

Second, we measure the change in consumption pace of the existing consumers in response to reduced wait-times. Specifically, we analyze the impact on excess wait-time, or how much consumers waited in excess of what is required. Given that the consumer's decision to access the free episode now versus later depends on the marginal consumption utility between two periods, the rate of complementarity decay over time dictates the direction in which the excess wait-time shifts due to reduced wait-times. The analysis reveals a 21% decrease in excess wait-time, suggesting that the complementarity value indeed decays over time and does so at a decreasing rate. 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 measure the impact of wait-time reduction on the inflow of new consumers into the series (extensive margin). As the a priori value of a given series increases under reduced wait-times, it incentivizes more consumers to start consuming the series for the first time. The estimation results report a 28% increase.

Finally, we bring together the individual pieces to examine the net effect of the wait-time reduction. We find that the reduction leads to a 92% and a 19% increase in aggregate consumption and purchases, respectively. Hence, despite the risk of cannibalization, the shorter wait-times actually increase platform revenues by stimulating paid consumption from existing consumers and attracting new consumers to the series. A back-of-the-envelope calculation indicates that the increase in intensive margin is responsible for 40% of the total increase in aggregate purchases. The estimated elasticity of consumption and purchases with respect to wait-times is 0.26% and 0.06%, respectively.

We then conduct a battery of robustness checks. To show that our estimated impact of wait-time reduction is indeed causal and unbiased, we conduct formal tests to confirm the parallel trends and SUTVA assumptions. We also conduct a sensitivity analysis that tests how strongly potential unobserved confounders would need to be associated with both the outcome and the treatment in order to explain away the estimated treatment effect. Moreover, to rule out spurious correlations that might drive our results, we conduct falsification tests using pseudo treatment series and dates. Analyses using different subsamples of the data and model specifications also show broad agreement with our main results.

The rest of the article is organized as follows. First, we discuss how this research is related to the literature. Second, we describe institutional details and data. Third, we explain the empirical strategy and the econometric model. Fourth, we discuss our conceptual framework based on the complementarity properties of serialized media. Fifth, we present the 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 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 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 investigate 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). In many of these cases, the key determinant of the firm's decision is the trade-off between consumer acquisition and cannibalization: an attractive free offering expands the consumer base, but does so at the cost of lower purchase per consumer. Our study enriches this dynamic by focusing on the unique complementarity properties of serialized media, which surprisingly dominates the negative cannibalization effect and encourages higher paid consumption per consumer when the free option is made more attractive.

Moreover, there is limited research that exploits changes in version quality to empirically investigate the causal effects, because product quality is difficult to quantify and a discrete change in version quality are uncommon. An exception is Li et al. (2019), who exogenously vary the resolution of free e-book 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 are poor substitutes. In contrast, 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 media, 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 open a secondary channel for mass distribution that sells at a lower margin (e.g., 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 (Calzada and Valletti, 2012; August et al., 2015; Luan and Sudhir, 2022). The literature focuses on cross-sectional discrimination assuming a static WTP for a product within consumer. While the wait-for-free policy for serialized media shares commonalities, it also allows for intertemporal discrimination within a consumer. Consumer's WTP for an episode varies across time as a function of complementarity from the previous episode and the value of her outside option. Hence, the platform can continuously discriminate between a consumer's states over the wait-time horizon.

With the proliferation of serialized media content, a growing stream of literature is giving attention to their monetization strategies. Zhao et al. (2022) investigates the connection between episode release timing and a set of consumer behaviors such as binge consumption, rationing and platform visits to study the optimal release schedule. By estimating a model of purchase and consumption of chapters within a book, the authors find evidence of both within-period binge consumption (i.e., conditional on availability, the consumer will consume a large number of chapters within a period) and intertemporal binge consumption (i.e., consumption in the last period begets consumption in the present period). These findings are consistent with our proposed complementarity properties of serialized media. The authors conclude that a hybrid strategy of simultaneous and sequential release strategies yield highest platform profits. Similarly, Godinho de Matos et al. (2023) devise an analytical model of consumer search and consumption on a TV show streaming platform, and conduct a field experiment to show that a drip-style content release schedule leads to higher platform usage.

There is limited research around the use of WFF policy on serialized media due to its novelty. The paper closest to ours is Choi et al. (2022), where they study the impact of introducing WFF to serialized comics that were previously pay-only. The authors find that the introduction resulted in higher paid episode views and mentions that the platform faces an implicit trade-off in revenues between cannibalization and incremental purchases from new users. Our work complements the study by exploiting an exogenous variation in wait-times to separately assess the impact on consumption depth, pace and the decision to start a series. Importantly, we draw on the foundational characteristics of serialization to hypothesize and empirically validate that lowering wait-times can actually lead to greater monetization even in the absence of purchases from new users. Another related study is Choi et al. (2023). Using data from a comics platform that allows early episode access for a fee, the authors find that greater availability of free episodes lead to habit formation, which in turn increases consumers' WTP for early access through payments. However, the WFF policy in their setting is similar to the static sequential release (Luan and Sudhir, 2022), and the model is focused on the purchase decision of the first non-free episode.

Finally, another related stream focuses on consumption behaviors of serialized media. Zhang et al. (2022) provides evidence of time-inconsistent preferences, where consumers intentionally choose to overpay for content in order to curb future consumption (strategic self-control). Several works study the phenomenon of binge consumption, exploring the implications on downstream behaviors such as responsiveness to advertisements, series completion and spillovers to other content on the platform (Schweidel and Moe, 2016; Lu et al., 2019, 2023; Godinho de Matos and Ferreira, 2020). These papers focus on providing empirical evidence of binge consumption and abstract away from the mechanism that drives binge consumption. We add to this stream of literature by shedding light on the role of serialization that affect consumption decisions.

3 Institutional Details and Data

3.1 Institutional Details

The serialized 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 20,000 series and has over a million active users. The platform generates revenues through users' episode purchases. Specifically, each series belongs to one of three categories depending on the monetization type: free, premium and wait-for-free (WFF). Free series allow immediate access to all episodes at no cost. Premium series follow a pay-per-episode model where first several episodes are free, and users must pay using an in-app currency ("Coins") 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. Like premium series, readers may alternatively pay to unlock the episode immediately. Coins can be bought with real money, and each episode costs 3 Coins regardless of wait-times, roughly equivalent to 50 cents. Users can earn Coins through other ways such as referrals, watching ads or giveaway events hosted by the platform, but those make up a negligible portion.

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* (e.g., a consumer returning in 12 hours will still 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 static release schedules. Also, note that there is no incentive for a user to purchase now to read later (i.e., stockpile), since the episode will eventually become free. To be clear, although firms have previously discriminated using time in contexts such as hardcover versus softcover books, the application to serialized media, where there are complementarities across episodes, has not been previously observed. Moreover, 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 WFF series constitute a third of all series on the platform, more than 85% of episode consumption in our dataset is generated by the WFF series. Given this pattern and our research objective, we focus only on the WFF series within the data.

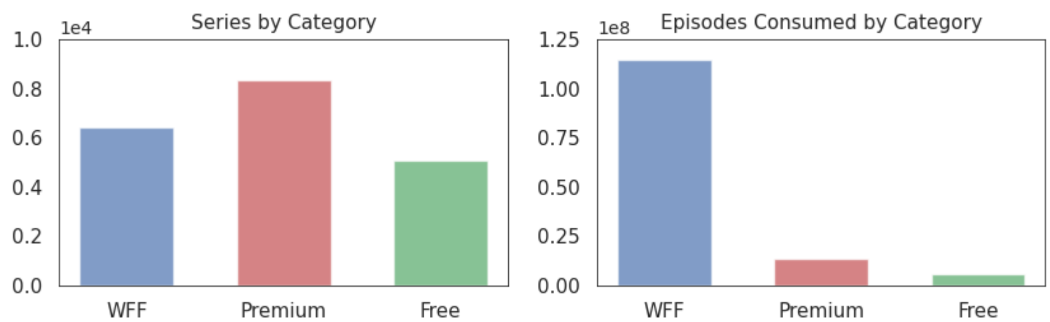

Figure 1: Distribution of series and episodes consumed across categories

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. The example below is a contemporary 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. A typical episode on the platform is around 1,500 words, and the vast majority of the readers finish an episode within 15 minutes.

3.2 Data

We leverage multiple datasets that cover user consumption as well as series and episode metadata. The consumption panel data covers 15 months from October 1, 2020 to December 31, 2021 and details when and how (waited or purchased) the user consumed the episode. Series metadata include title, genre, author, sales type, date of first publication and the required wait-time. Episode metadata include series ID, sequence in

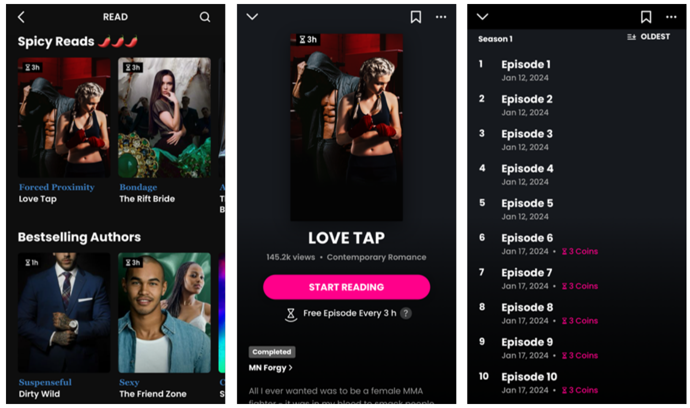

Figure 2: App user interface

the series, episode publication date and word count. The dataset also contains information on promotional activities where the platform offered coupons for specific series that can be used to unlock an episode, including the promotion dates and how many coupons were used. Finally, we have access to Coin purchase panel data from January 1, 2019 to October 30, 2022, which we leverage to explore heterogeneity across users based on historical spending. Our expansive dataset comprised of precise access timing and method over an extended time window presents a unique opportunity to delve into the consumption dynamics.

To isolate the effect of wait-time reduction on the *existing user base of the platform*, we filter the panel data to the users that joined the platform before October 1, 2020, the beginning of our consumption panel data. This ensures that there are no compositional changes to the platform population that might otherwise introduce a form of selection bias. This also means that our analysis results are conservative estimates that exclude the contribution from users that join the platform at a later date. Moreover, to reduce noise from tail end series that are rarely read, we filter for series with at least 1,000 episode accesses over the entire observation period. Our resulting dataset covers 1,940 WFF series and 308,681 users. The basic summary statistics about user consumption are provided in Table 1. The median series contains 44 episodes, and the median user has read two series and 57 episodes during the observation period. The median number of days that a user spends on a series is five. In general, the data is skewed to the right (with many heavy users) as the mean is usually much larger than the median, which we address in the analysis through log-transformation.

	Mean	**SD**	**25%**	**50%**	**75%**
Episodes per series	79.1	163.2	31.0	44.0	83.0
Series consumed per user	12.2	39.6	1.0	2.0	6.0
Episodes consumed per user	457.5	1487.9	10.0	57.0	267.0
Episodes waited per user	312.8	1233.7	2.0	18.0	141.0
Episodes purchased per user	132.4	392.3	1.0	16.0	87.0
Days spent per user per series	37.2	78.1	0.0	5.0	29.0

Table 1: Summary statistics on user consumption

We next provide a set of descriptive statistics around the series. The left panel of Figure 3 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 log-normal 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 total spending on the platform to purchase Coins. The most notable pattern is that there is a segment of users that do not spend any money on the platform (i.e., "non-spenders") that make up 53% of the user base. These users do not make any direct contributions to platform revenues and mostly resort to free consumption. We note that although the non-spenders do not purchase Coins, they may still purchase episodes using Coins earned through alternative methods mentioned above (e.g., referrals, ads, giveaway events). The remaining 47% of the user base show a log-normal distribution, indicating a long tail of heavy spenders. One could speculate that wait-times might affect consumers differently based on inclination to spend money, and the platform's primary concern is on the consumption decisions of the spenders who directly contribute to revenues, which calls the need to examine heterogeneity in the effect of wait-time changes across consumers based on historical spending in the empirical analysis.

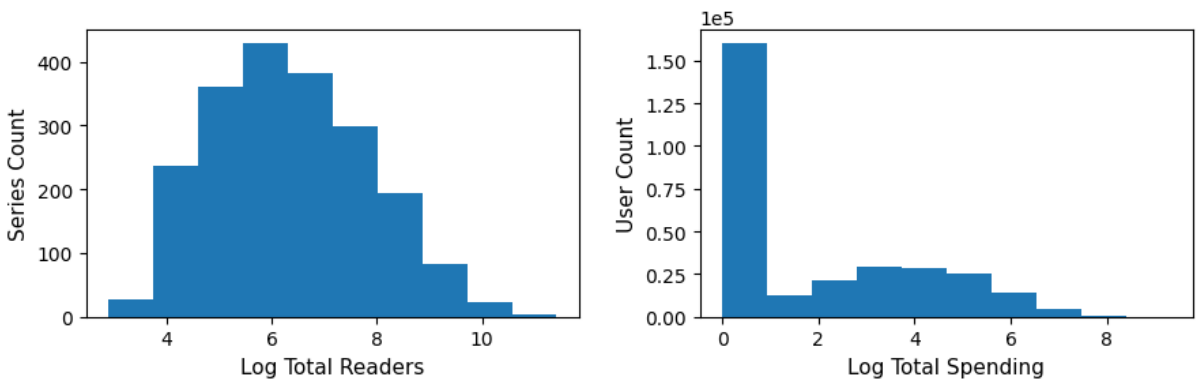

Figure 3: Distribution of readers across series and purchase propensity

Next, we explore how consuming an episode impacts the likelihood of the consuming the next episode in the series. Sampling the consumption panel data for 10,000 randomly selected users, we find that 99% of the episodes are read along with the immediately preceding episode, and 90% of the episodes are read in sequential order. A logistic regression reveals that reading an episode significantly increases the likelihood of reading the next episode ($p < 0.001$). In other words, the vast majority of readers read episodes in the specified order and read an episode only if they have read the preceding one.

We conclude this section by exploring consumption patterns across reading sessions. 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, we can infer that users visit the platform throughout the day mostly to access the waited episode and occasionally purchase another. 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. These patterns lend empirical support to the notion of directed complementarity between episodes that diminishes over time.

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 series read, conditional on the number of episodes consumed 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 and purchase decisions. An ideal experiment would randomly assign readers to different wait-times for a given series and compare outcomes between experimental designs. However, our setting features the same wait-time for all readers for a given series. Hence we leverage exogenous changes in wait-times implemented by the platform. Specifically, the platform reduced wait-times for a 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 for free after the reduction. While the overt reason for the policy change was to increase overall reader engagement, the series chosen for change, the revised wait times and the timing of the changes were not based on any systematic criteria. Importantly, the platform made no prior announcements about the changes; hence the changes can be treated as exogenous to readers. Similar instances can be found on many platforms where they unexpectedly implemented changes to 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 the dataset, which we call *treated series*. The rest of the 1,749 series did not have any changes to wait-times, which we call *non-treated series*. Figure 4 shows the distribution of the WFF series before and after treatment. As the wait-times were reduced, we see an increase in the proportion of series with 1-hour wait-times post-treatment and a corresponding decrease in the proportion of series with longer wait-times. A detailed breakdown of wait-times pre- and post-treatment is presented in Table 3: the diagonal and off-diagonal represent the number of non-treated and treated series, respectively. Figure 5 provides additional information about the policy change, including the distribution of treated series by adoption timing (left panel) and magnitude of reduction in hours (right panel).

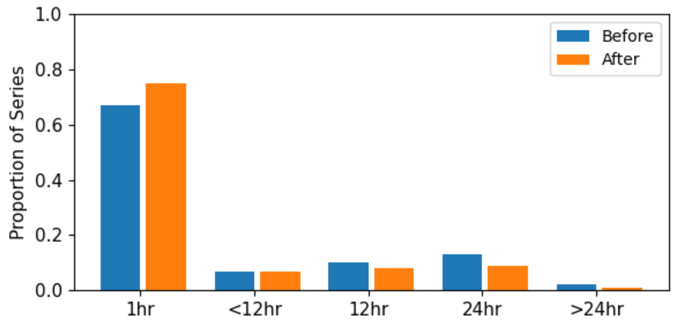

Figure 4: Distribution of series by required wait-time before and after treatment

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

We provide model-free evidence illustrating the impact of wait-time reduction on consumption patterns of existing and new consumers to the series. The left panel of Figure 6 compares the number of episodes purchased by an existing consumer of the series before and after the reduction for ten randomly sampled series. The plots indicate a noticeable increase in paid consumption under reduced wait-times. The middle panel shows the excess wait-time (i.e., how much time (hrs) the existing consumers waited to access the free episodes in excess of what is required) for the same set of series. The downward sloping lines suggest

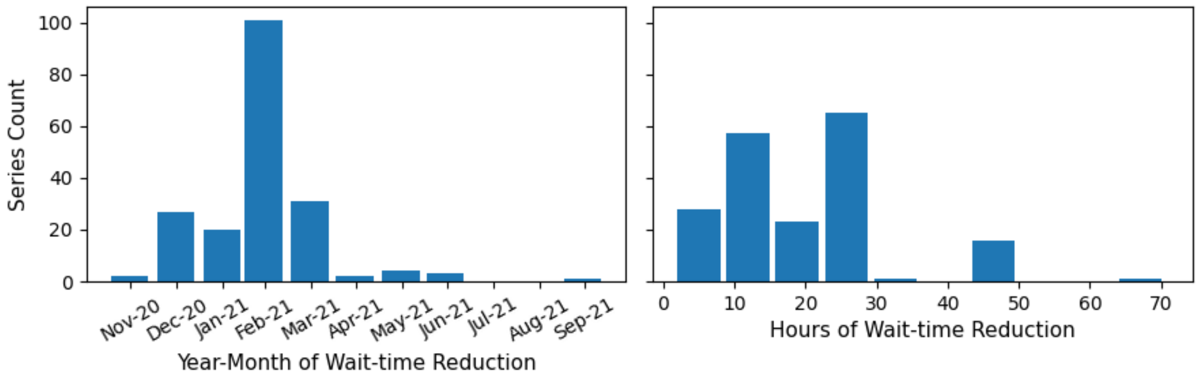

Figure 5: Distribution of treated series by treatment timing and reduction magnitude

that consumers wait less in excess, or in other words, consume at a faster pace. The right panel displays the number of unique consumers for a given series for the same set of series. The positive slopes indicate an expanding consumer base under shorter wait-times, potentially generating incremental revenues for the platform. For all three measures, a paired t-test at the consumer-level across the entire set of treated series demonstrates that the values before and after the reduction are significantly different ($p < 0.001$).

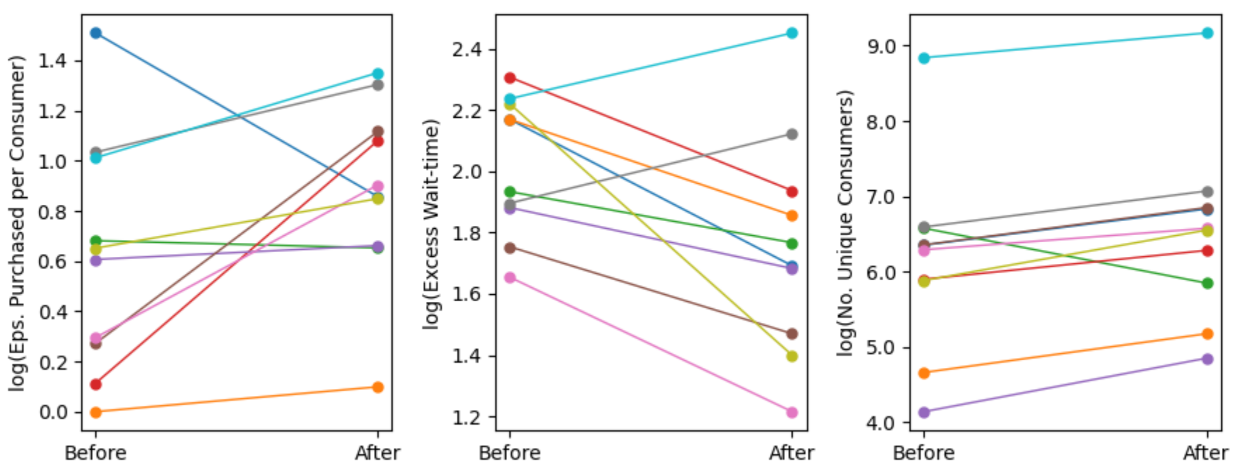

Figure 6: Comparison of episodes purchased (left), excess wait-time (middle) and the number of unique consumers (right) for ten randomly selected treated series before and after the reduction

Alternatively, we visualize how retention across episodes in a given series changes upon the wait-time reduction. Figure 7 illustrates the average proportion of consumers that consume the first episode either before or after the reduction that proceed to consume subsequent episodes of the treated series. The plots overall indicate high churn especially in the early episodes, and less than 40% of the readers remain by the 50$^{\text{th}}$ episode. Importantly, we see a noticeable increase in retention under reduced wait-times.[2]

[2]Note that there are two potential explanations for such increase: each consumer may be progressing further in the episode (increase in intensive margins), or the population of readers starting the series may be different before and after the reduction.

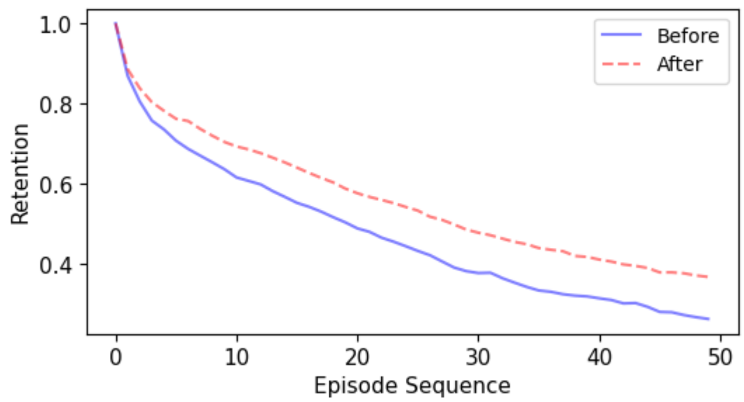

Figure 7: Average retention across episodes

Together, Figures 6 and 7 lend empirical support to our theoretical arguments: as the consumer waits less for free access, she receives a higher complementarity value and thus decides to consume the episode rather than churn, which in turn allows her to benefit from complementarity with the next episode. Such increase in retention may result in increased purchases despite the higher incentives to wait for free. We next proceed to construct a robust causal model that considers the impact of wait-time reduction on consumption while controlling for a host of observable and unobservable features.

4.1.1 Empirical Challenges

To identify the effect of shortening the wait-time on individual consumption and aggregate demand, we focus our analysis on a 30-day 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. By comparing the treated series to an appropriately constructed set of control series with no wait-time changes, we can estimate the effect using a difference-in-difference (DiD) framework.

As mentioned earlier, our empirical context poses two main challenges for such a DiD analysis. The first is the selection into treatment. Although the platform indicated that they did not use specific selection criteria for wait time reduction, the treated and non-treated series could still be systematically different, potentially leading to biased results if we did a naive comparison. For example, if the treated series had previously been more widely read, then the estimated treatment effect will be downward biased, as the wait-time reduction has no effect on users that have already read the series. If the platform had selected longer series for reducing wait times, then the estimate would be upward biased, as more episodes are affected by the reduction.

The second challenge is that we have unbalanced panel data with variation in treatment timing. Since series are published on or removed from the platform at different points in time, the observed time window

We later isolate the impact on the intensive margins by focusing the analysis on the existing consumers of the series.

varies across series (only about 6% of the series are removed during the observation period). 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; Sun and Abraham, 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 propensity score, which is the probability of receiving the treatment conditional on the pre-treatment covariate histories. Hence, treatment assignment is independent of potential outcomes conditional on potential outcomes, satisfying the conditional independence assumption. Any difference in observed outcomes between the treated and matched control units can be attributed to the treatment effect, thereby approximating a randomized experimental design.

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

4.2 Constructing a Matched Control Set via Panel Matching

In order to address the potential systematic differences between the treated and the non-treated series, we create a control group for each treated series by matching it with non-treated series that have similar probability of being treated. By making treatment independent of observed potential confounders (i.e., conditional independence assumption: $Y(0), Y(1)) \perp\!\!\!\perp T|X)$, we can draw causal conclusions about the impact of reduced wait-time by comparing the two groups. However, most existing applications with matching assume a cross sectional dataset using static features measured at a point in time (Abadie and Imbens, 2011; Diamond and Sekhon, 2013; Hansen, 2004). When applications use panel data, they compute the average of time-varying covariates over a static time-frame (Datta et al., 2018; Deng et al., 2022; Narang and Shankar, 2019) to fit the cross-sectional matching framework. However, this can miss out on important time-varying factors such as demand trends leading up to treatment that affect selection into treatment.

In our setting, some of the potential demand-related confounders (e.g., number of episodes waited and purchased) are time-varying, and the variation in treatment timing makes it difficult to define a single pre-treatment period for the non-treated series. Moreover, matching on the average of time-varying covariates might match series whose covariates are similar on average but exhibit very different temporal trajectories. For example, a series that is gaining traction among readers and one that is becoming increasingly unpopular prior to treatment will clearly experience different effects from reduction in wait-time. We therefore adapt the propensity score matching procedure for time-series cross-section data (panel-matching) developed in Imai et al. (2021) to address the issue of time varying factors.

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. As an illustrative example, Figure 8 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 fully observed (blue areas) around the treatment timing and are comparable in covariate values leading up to treatment.

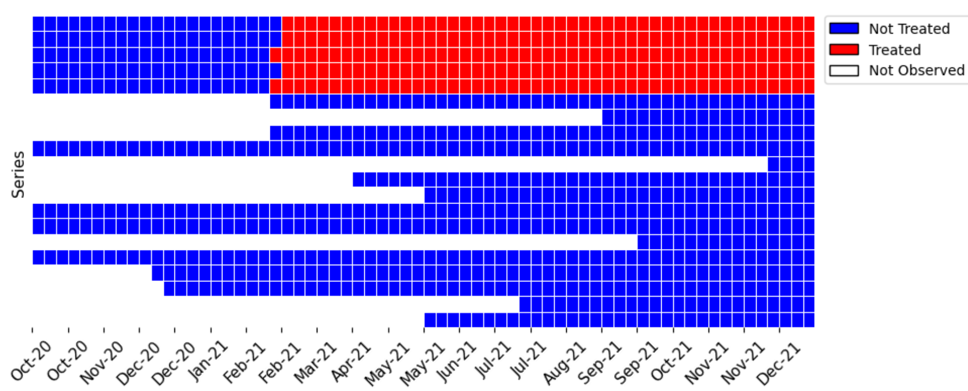

Figure 8: Treatment Variation Plot

Constructing the Matched Control Set We now describe the matching procedure in detail. Let us denote a treated series s that receives treatment for the first time 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 $t-1$.[3] Figure 9 illustrates an example of how matched sets are created when $L=2$, indicated by the color of the boxes around the observations. In this example, treated series $s=0$ is matched to non-treated series $s \in \{3,4\}$ over weeks $t \in \{0,1\}$ (blue box). Note that these non-treated series are fully observed in the two weeks prior to the treatment timing of series $s=0$. Series $s \in \{1,2,5\}$ are not included in the matched set because they are either eventually treated or are not fully observed. Similarly, series $s=1$ is matched to series $s \in \{4,5\}$ (red box), and series $s=2$ is matched to series $s \in \{3,4,5\}$ (green box). In our case, we set $L=4$, which assumes that adjusting for covariate trends up to previous four weeks 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 9: Illustrative example of constructing the matched control set. The color of the boxes indicate the treated and matched control units included in the same set. For example, treated series $s=0$ is matched to non-treated series $s \in \{3,4\}$ over weeks $t \in \{0,1\}$ (blue box).

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 pre-treatment covariates (Rosenbaum and Rubin, 1983). The propensity score is computed using a logistic regression based on a set of observed covariate histories prior to treatment that can reasonably discriminate the treated and non-treated series:

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

[3]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.

where $V_{s,t}$ is a matrix of observed static and time-varying covariates for series s in week t. The covariates used in the logistic regression include weekly count of waited/purchased episodes, series length, weeks since the series was first published and the required wait-time. The use of endogenous pre-treatment variables (i.e., waited and purchased episodes) to compute propensity score is consistent with the existing research that utilize covariates such as lagged outcomes, consumer spending and income (Heckman et al., 1998; Dehejia and Wahba, 2002). These covariates serve as critical proxies for latent variables that might influence both the selection into treatment and the post-treatment outcomes of interest. By incorporating these variables, we are able to indirectly adjust for unobservable confounders and satisfy the parallel trends assumption.

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. Because the treatment timing varies across the treated series, potential concerns about over-reliance on specific control units from matching with replacement are mitigated. 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 10 compares the covariate balance across $N \in \{1, \ldots 10\}$ (x-axis) based on p-values of the t-test (left y-axis) and Kullback–Leibler divergence (right y-axis). The blue line indicates the number of covariates for which the p-value is greater than 0.01. The plot shows good balance for $N \leq 7$, beyond which the matched series become increasingly different from the treated series in terms of covariate means. The green dotted line indicates KL divergence of the propensity scores. KLD closer to zero indicate greater similarity between two distributions; the balance across the two groups are stable across N.

Matching a treated unit to a single or multiple control units are both widely used in practice, each approach having its own precision-bias tradeoff. One-to-one matching typically yields less bias because each treated unit is matched with its closest counterpart, but it results in fewer matches and thus less precision. On the other hand, 1:N matching can enhance precision by utilizing more data (i.e., more comparisons), but may introduce bias if the matched units are not as similar to the treated unit, or if certain controls are

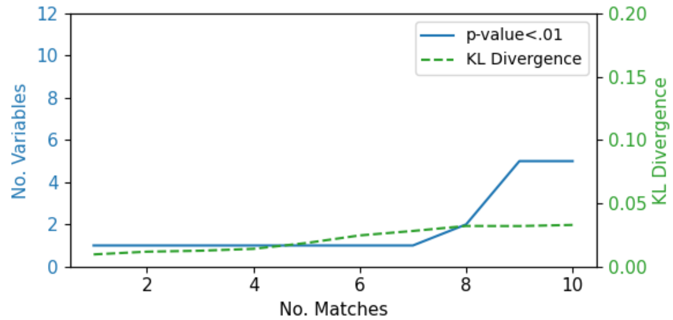

Figure 10: Assessment of match quality by the number of matches

overutilized. To balance obtaining a sufficiently large sample against the risk of potential overfitting, we proceed with $N = 5$, beyond which some treated series start to lack additional eligible matches within the caliper. By setting the difference in propensity scores of matched units to be less than a caliper of $C = 0.1$, we mitigate potential bias concerns. Additionally, we conduct robustness checks to ensure that the results remain consistent for $N \in \{1, \dots 10\}$.

Table 4 evaluates the balance of the covariates and propensity score distribution before and after matching at $N = 5$. The results show that the treated series and their matched control series are not significantly different in key variables based on p-values of the t-test with the exception of wait-times. We note that while it is challenging to achieve perfect balance in wait-times given the concentration of series with 1-hour wait-times, the matching process markedly reduces the gap compared to before matching. Figure 11 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.

	Treated	Control (Before Matching)		Control (After Matching)	
	Mean	Mean	p-value	Mean	p-value
Propensity Score	0.697	0.307	0.000	0.696	0.918
log(T1 waited + 1)	5.304	4.989	0.231	5.075	0.123
log(T2 waited + 1)	5.263	5.003	0.577	5.043	0.150
log(T3 waited + 1)	5.252	5.000	0.605	5.038	0.162
log(T4 waited + 1)	5.217	4.987	0.702	4.959	0.093
log(T1 purchased + 1)	4.595	4.006	0.000	4.499	0.524
log(T2 purchased + 1)	4.654	4.020	0.000	4.604	0.728
log(T3 purchased + 1)	4.508	4.015	0.000	4.440	0.666
log(T4 purchased + 1)	4.456	4.020	0.000	4.386	0.658
log(No. Episodes)	4.038	4.035	0.217	4.004	0.576
log(Weeks since published)	4.253	3.785	0.000	4.148	0.122
Wait-time (Standardized)	0.954	-0.012	0.000	1.547	0.000

Table 4: Covariate balance across treated and matched control series before and after matching

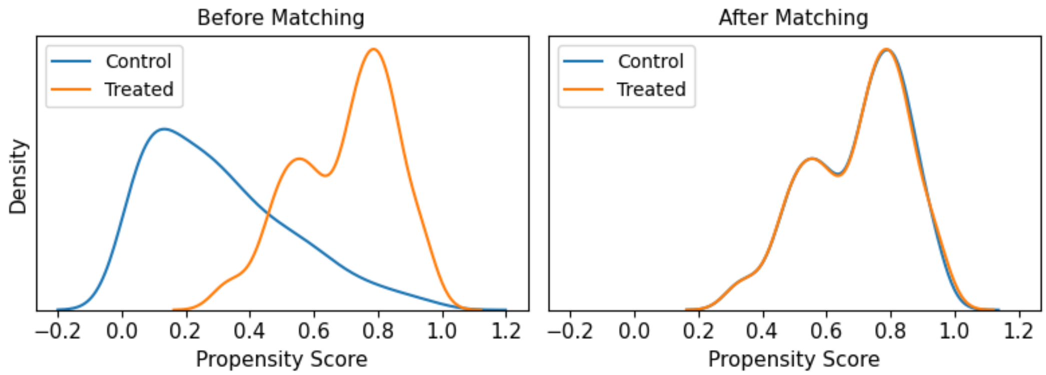

Figure 11: Propensity score distribution before and after matching

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 DiD 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. When treatment effects vary across both time and units, 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, which 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).

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 five matched control series, which we denote as a *series group* (or cohort) 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 period fixed effects, which fully controls for self-selection on unobserved time-invariant factors. Gardner (2022) shows that this approach estimates a convex weighted average of the individual treatment effects under parallel trends and no anticipation.

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 *Group-Series FE*); ν_{gp} is a fixed effect specific to group g in period p, which captures unobservable time trends (referred to as *Group-Period FE*). By including group specific fixed effects, the model essentially estimates the DiD from each series group and then applies variance weighting to aggregate the treatment effects (Baker et al., 2022). Finally, ϵ_{sgp} is the error term, which are clustered at the series level.

5 Conceptual Framework

In this section, we propose a conceptual framework that forms our expectations around the impact of changing wait-times on consumption and purchase behaviors before embarking on the empirical analysis. As mentioned earlier, our conceptual framework leverages three unique properties of serialized media. First, episodes in a series share structural and content similarities that in tandem with literary devices such as cliffhangers give rise to directed complementarities. Second, the complementarity value decays over the consumption interval as consumption capital dissipates in time. Third, the consumer's willingness-to-pay for a given episode exogenously varies over time based on the value of the consumer's outside option.

Consider a consumer making consumption decisions for episode e in a series with wait-time T hours. In each period t, the consumer may choose among actions $a \in \{consume, wait, exit\}$ for the episode. The consumer's action yields *episode flow utility*, u_e^a, defined as follows:

$$u_e^a(t,\tau,T) = \begin{cases} v_e + C(t,\tau) - P(t,\tau,T) + \epsilon_t^a & \text{if } a = consume \\ \epsilon_t^a & \text{if } a \in \{wait, exit\} \end{cases} \tag{5}$$

where v_e is the baseline valuation for episode e, $C(t,\tau)$ is the complementarity value for the episode at period t from having consumed the previous episode in period τ, $P(t,\tau,T)$ is the episode price at period t given τ and wait-time T, and ϵ_t^a is an exogenous period-specific Type I extreme value shock. Her complementarity value for the episode decays over the time since consuming the previous episode (i.e., C is a decreasing function in $t-\tau$). With the WFF option, she may consume the episode by paying before the wait-time elapses or for free T hours after consuming the previous episode (i.e., $P = p$ if $t-\tau < T$; $P = 0$ if $t-\tau \geq T$).

The consumer considers each episode in order and can only consume if she has consumed the previous episode.[4] If she consumes episode e in period t, she proceeds to consider purchasing episode $e+1$ in the same period; if she waits, she considers the same decision for episode e in the next period $t+1$; if she exits,

[4]Data patterns in Section 3.2 show that consumers typically read an episode only if they have read the preceding episode.

she no longer considers any subsequent episodes.

We further make the following assumptions. First, the consumer's baseline valuation for the next episode v_{e+1} is only realized after consuming episode e, but she knows the distribution from which it is drawn. This reflects the realistic scenario where the consumer may have expectations about episode quality based on genre or plot, but is uncertain until she experiences the content of the preceding episode.[5] Second, the consumer is rational and forward looking, and thus accounts for the fact that consuming episode e allows her to proceed to episode $e + 1$. Third, consumption is concurrent with purchase, since there is no incentive for the consumer to stockpile (i.e., purchase now to consume later), as noted in Section 3.1.

Given the above setup, the consumer consumes episode e in period t after evaluating utilities based on expectations over the utility across all periods. Actions *consume* and *wait* allow the consumer to proceed to the next episode, and hence, provide *inclusive value*, or the expected maximum utility from her optimal decisions made onwards. Action *exit* does not provide inclusive value as the consumer does not consume any more episodes. Formally, the consumer chooses an action by solving the following value function:

$$V_e(t, \tau, T) = \max\{U_e^{consume}(t, \tau, T), U_e^{wait}(t, \tau, T), U_e^{exit}(t, \tau, T)\} \tag{6}$$

where

$$U_e^{consume}(t, \tau, T) = u_e^{consume}(t, \tau, T) + V_{e+1}(t, t, T) \tag{7}$$

$$U_e^{wait}(t, \tau, T) = u_e^{wait}(t, \tau, T) + \beta V_e(t + 1, \tau, T) \tag{8}$$

$$U_e^{exit}(t, \tau, T) = u_e^{exit}(t, \tau, T) \tag{9}$$

We illustrate the potential implications of reducing wait-times by simulating the consumption decisions for episodes in a series with three episodes, $e \in \{0, 1, 2\}$, under wait-times $T = 2$ and $T = 1$. The wait-time ticks down conditional on consuming the previous episode, so we begin with her consumption decision for episode 1. For simplicity, we simulate decisions over five periods and assume that if the consumer has not consumed an episode over three periods, she no longer considers the episode (i.e., churns from the series). Functional form and parameter assumptions used for the simulation are presented in Table 5. Since we consider a finite sequence of episodes, the inclusive values can be easily computed via backward induction.

Table 6 shows the decision matrix of 1,000 simulations. The values indicate the number of consumers for an action pair in each wait-time regime, reflecting the impact solely due to the wait-time reduction. The left table shows the decisions for episode 1. First, a consumer that paid under $T = 2$ may now choose to consume for free under $T = 1$ (64). She now gains a higher flow utility from free consumption, thereby reducing the

[5] While our model allows v_e to vary across episodes, one may simply assume v_e is fixed across episodes.

	Assumption
Baseline valuation	$v_e \sim \mathcal{N}(0,1)$
Complementarity value	$C_e(t,\tau) = 7 \cdot e^{-2(t-\tau)}$
Initial price	$p = 7$
Period specific shock	$\epsilon_t \sim Gumbel(0,1)$
Discount factor	$\beta = 0.9$

Table 5: Functional form and parameter assumptions for the simulation

likelihood that purchasing yields the highest expected net utility among the actions in any given period. This represents the cannibalization effect widely documented in the freemium literature (Lee et al., 2019; Li et al., 2019; Cao et al., 2023). At the same time, a consumer that chose exit (i.e., no consumption) or free consumption under $T = 2$ may now choose to consume for free (275) or pay (32), because both the flow utility and the inclusive value from those actions are greater. We also see consumers switch from pay to exit (15) due to the shift in consumption timing and the exogenous period specific shocks that follow.

The right table shows the impact for episode 2. Again, we see purchase cannibalization (40). If the consumer did not consume episode 1, she would not consume episode 2, as seen in the higher total exit counts in episode 2 compared to episode 1. On the flip side, those that switched from exit to consumption in episode 1 can now consider episode 2 with the benefit of complementarities. If the realized valuation is sufficiently high, she chooses to purchase (95). Again, we see consumers switching in the below-diagonal due to the shifts in consumption timing. The impact on the intensive margins of a given series is the sum of changes changes across all episodes in the series. Note that this does not include the positive impact of incremental purchases from new consumers (i.e., expansion of extensive margins).

Decision Matrix for Episode 1

2hr/1hr	Exit	Pay	Free	Total
Exit	258	32	275	565
	45%	6%	49%	100%
Pay	15	285	64	364
	4%	78%	18%	100%
Free	-	-	71	71
	-	-	100%	100%
Total	273	317	410	1000

Decision Matrix for Episode 2

2hr/1hr	Exit	Pay	Free	Total
Exit	547	95	161	803
	68%	12%	20%	100%
Pay	24	103	40	167
	14%	62%	24%	100%
Free	3	6	21	30
	10%	20%	70%	100%
Total	574	204	222	1000

Table 6: Decision matrices for episodes 1 and 2 under 1-hour and 2-hour wait-times

The change in the intensive margin from reduced wait-time is thus the net impact of the negative cannibalization effect and the positive across-episode spillovers arising from complementarities. We expect to find an increase in overall consumption, but whether paid consumption will increase remains an empirical question, as it is ex-ante unclear which of the two effects would dominate. Importantly, note that in the absence of the complementarity properties, purchases should *always* decrease. In this case, the consumer's WTP for an episode would vary over time only depending on the exogenous shock, and there is a higher chance that her WTP is greater than the price in any period before the episode becomes free.

Next, the diminishing complementarity value over time also has implications on the consumption pace of the existing consumers. Specifically, the impact of wait-time reduction on how quickly consumers consume waited episodes would differ based on the rate of complementarity decay over time. Consider a consumer deciding when to consume a free episode after the wait-time has elapsed. Under wait-time $T = 2$, her decision to consume now in period $t = 2$ versus deferring to period $t = 3$ hinges on the expected change in utility between these periods. Similarly, with a halved wait-time of $T = 1$, the decision depends on the expected change in utility between periods $t = 1$ and $t = 2$.

If complementarity diminishes in a linear fashion over time, there should be no change in how much consumers wait in excess for a free episode (i.e., excess wait-time), because the marginal utility between the successive periods is the same under both wait-time regimes. If the complementarity value diminishes at a decreasing (increasing) rate, the excess wait-time should decrease (increase), as deferring consumption has a larger (smaller) negative impact under shorter wait-times. The convex trajectory of the complementarity value aligns with the extensive literature on forgetting, which shows that consumer recall exhibits a rapid initial decline right after an event, followed by a slow decay (Clarke, 1976; Mahajan and Muller, 1986). Conversely, the concave trajectory may be justified by a gradual information overload over time (Klingberg, 2009). As the consumer engages in more outside content after consuming an episode, the cognitive effort required to reconnect with the series increases, and hence the perceived value of re-engaging may decline more steeply. There is no inherent reason to believe that complementarities in serialized media would exhibit any specific trajectory, making this an empirical question. The pace of free consumption is critical to the platform, as it induces the consumer to make consumption decisions on the following episodes.

Finally, the expansion of the consumer base from increased value of the free option is well-understood from the freemium literature. For serialized media, a consumer decides to start a series (i.e., consume the first episode) if the aggregate expected utility from the episodes outweighs the cost. It is clear from the above setup that the aggregate expected utility is higher under wait-time $T = 1$ due to higher flow utility and inclusive value. Hence, we expect growth along the extensive margin.

6 Analysis and Results

Our analysis aims to measure the impact of changing wait-times on consumption decisions and monetization of serialized media. The empirical analysis proceeds in four steps. First, we examine the effect of wait-time reduction on the intensive margin, or the consumption depth of the existing consumers of the series. This enables us to test for the existence of positive across-episode spillover and empirically measure the impact net of the negative cannibalization effect. Second, we examine the change in the consumption pace of the existing consumers in response to the reduction. The results would inform us about the shape of the complementarity decay. Third, we examine the effect on the extensive margin, or the decision to start a series by measuring the inflow of consumers who consume the series for the first time. Finally, we measure the overall effect of reduction on aggregate consumption and revenues, complemented with a battery of robustness checks.

6.1 Intensive Margin: Depth of Series Consumption

Conditional on starting a series, how far does a consumer progress and how many episodes does she purchase along the way? For serialized media, the complementarity from having consumed the previous episode represents a significant portion of her valuation for the episode. Hence, if a consumer decides not to consume an episode, the platform loses out on the sales of not only that episode, but also all potential sales on subsequent episodes. Our conceptual framework illustrates how reducing wait-times may serve as a solution to increase purchases from the existing consumers, despite the risk of cannibalization.

We begin by sampling consumers that consume episodes both within 15 days before and after treatment for each treated and control series. We measure the impact of wait-time reduction on the total number of episodes consumed (i.e., waited and purchased) and separately for episodes purchased by running the stacked DiD regression from Equation 4 with the two-period consumer panel data. The outcome variables are log-transformed to address the skewed distribution of the data. The model controls for observable characteristics that may affect consumption, including series length, days since first and last episode publication and the presence of promotions. In addition to the group-series and group-period fixed effects, we also include individual fixed effects based on pre-treatment spending behavior. Specifically, we categorize the consumers into two types based on whether they had ever purchased Coins on the platform prior to the earliest treatment date. The stacked dataset consists of 13,405 spenders (44%) and 16,742 non-spenders (56%).

Estimation results are presented in column (1) of Table 7. The negative and significant coefficient on *non-spender* suggests that consumers inclined to spend money on the platform consume more episodes. The treatment effect of interest is the coefficient of *after* × *treated*. We find that the reduction of wait-times lead to a significant increase in episodes consumed, meaning consumers on average progress further in the series. The direction of the result is expected as shorter wait-times encourage consumers to consume episodes who might have otherwise chosen not to. To assess the magnitude of the treatment effect, denote the consumption count before and after the reduction as N_0 and N_1, respectively. Using the estimated coefficient $\hat{\beta}$, we can compute $N_1 = e^{\hat{\beta}}(N_0 + 1)$ and the percentage change in the dependent variable as $(N_1 - N_0)/N_0$. The estimate 0.289 in column (1) suggests that if the episodes consumed per series per consumer before the reduction is at the mean (9.1), it would increase to 12.5, a 37% increase.

Next, we estimate the impact of wait-time reduction on the total number of episodes purchased by a consumer for a given series. Increased consumption would only be detrimental to revenues if it came at the expense of less purchases. In the absence of across-episode complementarity that affect the consumer's willingness-to-pay, purchases should *always* decrease as the reduced wait-times would only yield cannibalization effect. Results are presented in column (3) of Table 7. The coefficient on *non-spender* is again negative and significant, meaning consumers inclined to spend money on the platform purchase more episodes.[6] More importantly, the estimated coefficient of *after* $\times$ *treated* is positive and significant at 0.073, indicating a 12% increase in paid consumption per consumer. This result shows that allowing quicker free consumption leads to greater depth in consumption *and* purchases. In other words, the platform can actually monetize the same consumer that may switch from purchasing to waiting for an episode to a greater extent.

In columns (2) and (4), we explore heterogeneity in treatment effect by allowing the effect to vary based on pre-treatment spending behavior. As seen in Figure 3, there is a segment of consumers unwilling to spend any money on the platform regardless of the wait-time, and hence, we would not expect to find a strategic shift in their purchase behavior from the wait-time reduction. Moreover, the platform is primarily concerned with the impact on spenders, because their baseline purchases are significantly higher than that of the non-spenders as denoted by the negative coefficient on *non-spender*, and any change in purchases of the non-spenders would have a minimal affect on revenues. The coefficient on *after* $\times$ *treated* represent the impact of wait-time reduction on spenders, and the coefficient on *after* $\times$ *treated* $\times$ *non-spender* separates out the incremental effect on the non-spenders. While we find no heterogeneity in the effect on total consumption, the effect on episodes purchased is magnified to 0.107, representing a 17% increase relative to the pre-treatment levels. Meanwhile, the effect on non-spenders is positive but no longer significant, meaning their purchase behavior does not significantly change, consistent with our expectations. In other words, while both segments of consumers consume more episodes, it is driven by increased waited episodes for non-spenders, whereas those inclined to spend money on the platform also purchase more episodes.

To assess robustness, we replicate the analysis with a 60-day window around the reduction. The results are qualitatively unchanged. Further, we estimate the model using the magnitude of wait-time reduction rather than a treatment indicator. That is, we replace *treated* with *magnitude*, computed as the log difference in wait-times pre- and post-treatment. Under our conceptual framework, the increase in complementarity

[6] Note that non-spenders have non-zero purchases because there are alternative ways of obtaining Coins other than spending real money, such as referrals, watching ads or giveaway events hosted by the platform.

	log(Consumed+1)		log(Purchased+1)	
	(1)	(2)	(3)	(4)
non-spender	-0.028***	-0.027***	-0.179***	-0.174***
	(0.010)	(0.010)	(0.023)	(0.025)
after × treated	0.289***	0.295***	0.073**	0.107***
	(0.031)	(0.035)	(0.029)	(0.035)
after × treated × non-spender		-0.014		-0.071**
		(0.023)		(0.032)
log(no. episodes)	-0.329	-0.331	-0.882***	-0.890***
	(0.341)	(0.341)	(0.336)	(0.336)
log(days since first pub.)	-0.982***	-0.982***	0.302**	0.302**
	(0.173)	(0.173)	(0.126)	(0.126)
log(days since last pub.)	-0.074***	-0.074***	-0.006	-0.007
	(0.023)	(0.023)	(0.012)	(0.012)
$\mathbb{1}$(promotion)	-0.044	-0.044	0.074**	0.073**
	(0.038)	(0.039)	(0.034)	(0.034)
Group-Series FE	Y	Y	Y	Y
Group-Period FE	Y	Y	Y	Y
N Obs	405634	405634	405634	405634
N Series Groups	191	191	191	191
R-squared Adj.	0.004	0.004	0.009	0.010

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

Table 7: Treatment effect on consumption and purchases per series per consumer

value and hence the treatment effect, should be proportional to the magnitude of the wait-time reduction. Indeed, the estimated coefficients on *after* × *magnitude* are positive and significant ($p < 0.01$).

6.2 Pace of Series Consumption

The WFF policy aims to establish regular and frequent user engagement by setting wait-times conditional on the consumption timing of the previous episode, so that the consumer periodically visits the series to claim the free episode and reset the clock for another free episode. Thus, it is important for the platform that the consumer accesses her free episode shortly after she becomes eligible. Our conceptual framework reveals that the direction in which the consumption pace shifts would depend on the trajectory of the diminishing complementarity value over time: a linear decay would lead to no changes in consumption pace beyond the mechanical shift from changing wait-times, whereas a decreasing (increasing) rate of decay would accelerate (decelerate) consumption pace due to differences in the consumer's marginal utility.

To empirically explore this dynamic, we investigate how wait-times affect *excess wait-time*. We define excess wait-time to be the time interval between when an episode becomes free for a consumer and when she actually consumes it. Low excess wait-time would indicate that the consumer consumed the episode soon after it became free. As before, we take the set of existing consumers within 15 days before and after treatment for each series and estimate the stacked DiD regression from Equation 4 with the two-period consumer panel data. The dependent variable is excess wait-time in hours (log-transformed). The model controls for observable characteristics that may affect consumption pace, including series length, relative position of the episode (episode number divided by series length), days since first and last publication and the presence of promotions. Also included are group-series and group-period fixed effects, as well as individual fixed effects based on pre-treatment spending behavior.

Table 8 reports the estimates. The coefficient on *after* $\times$ *treated* in column (1) reports a significant decrease in excess wait-time. If the excess wait-time before the reduction is at the mean (3.4 hours), it would decrease to 2.7 hours, a 21% decrease. The heterogeneous treatment effects estimated in column (2) report that the acceleration effect is even greater for non-spenders. The heterogeneity suggests that non-spenders may have a higher rate of decay in WTP during earlier periods compared to the spenders, leading them to wait than purchase. Hence, the marginal utility gain from accelerated consumption is greater. Our findings indicate that the declining complementarity value exhibits a convex trajectory with respect to time, consistent with the existing literature on forgetting. To confirm robustness, we use a wider window of 60 days around the treatment date and the alternative model specification using magnitude of wait-time reduction. In both cases, treatment effect estimates remain qualitatively unchanged ($p < 0.01$).

	log(Excess Wait-time + 1)	
	(1)	(2)
non-spender	-0.029***	-0.019*
	(0.010)	(0.011)
after $\times$ treated	-0.181***	-0.149***
	(0.046)	(0.046)
after $\times$ treated $\times$ non-spender		-0.061***
		(0.023)
episode position	0.582	0.583
	(0.434)	(0.434)
log(no. episodes)	2.762***	2.761***
	(1.043)	(1.042)
log(days since first pub.)	1.474***	1.473***
	(0.433)	(0.433)
log(days since last pub.)	-0.210***	-0.209***
	(0.057)	(0.057)
$\mathbb{1}$(promotion)	-0.072	-0.073
	(0.082)	(0.082)
Group-Series FE	Y	Y
Group-Period FE	Y	Y
N Obs	3091134	3091134
N Series Groups	187	187
R-squared Adj.	0.018	0.018

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

Table 8: Treatment effect on excess wait-time

6.3 Extensive Margin: Start of Series Consumption

Next, we investigate the impact of wait-times on the consumer's decision to consume a series for the first time. The reduction in wait-time increases the expected utility of the episodes and in turn the aggregate expected utility of the series. Hence, we expect an expansion in the extensive margin, with a larger stream of new consumers starting the series who would serve as sources of potential revenue.

To measure the impact on new consumer inflow at the series level, we estimate Equation 4 with a two-period series panel data during the 15 days before and after the reduction. The dependent variable is the number of consumers who read an episode of a series for the first time (log-transformed). The control variables include series length, days since first and last publication, and the presence of promotions. As the analysis is on a series level, we only include group-series and group-period fixed effects.

Results in column (1) of Table 9 demonstrate that the wait-time reduction leads to a significant increase in new consumers. At the pre-treatment mean of 96, the wait-time reduction increases the number of new consumers to 123, a 28% increase. Columns (2) and (3) report regression results separately for spenders and non-spenders. The treatment effects are positive and significant for both groups – at the pre-treatment mean, the number of new spenders increase from 47 to 63 (34% increase), and the number of new non-spenders increase from 49 to 60 (23% increase). The results are robust to a wider time window of 60 days around treatment and the alternative model specification using magnitude of wait-time reduction ($p < 0.01$).

	log(New Consumers + 1)		
	(1) All Consumers	(2) Spenders	(3) Non-spenders
after × treated	0.245***	0.285***	0.210**
	(0.082)	(0.084)	(0.092)
log(no. episodes)	-0.473	-0.204	-0.767
	(0.881)	(0.914)	(0.983)
log(days since first pub.)	-0.097	-0.105	-0.098
	(0.866)	(0.823)	(0.946)
log(days since last pub.)	-0.161	-0.189	-0.112
	(0.145)	(0.161)	(0.131)
$\mathbb{1}$(promotion)	1.341***	1.331***	1.333***
	(0.335)	(0.385)	(0.296)
Group-Series FE	Y	Y	Y
Group-Period FE	Y	Y	Y
N Obs	2292	2292	2292
N Series Groups	191	191	191
R-squared Adj.	0.122	0.127	0.092

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

Table 9: Treatment effect on inflow of new consumers

6.4 Aggregate Consumption and Purchases

Findings thus far provide empirical evidence to validate our conceptual framework. The existing consumers of the series purchase more episodes despite the risk of cannibalization and consume at a faster pace, accelerating potential purchases for subsequent episodes. Moreover, a greater stream of new consumers begin consuming the series, also contributing to revenues. However, the central question from the perspective of the firm remains: What is the net effect of the wait-time reduction on *aggregate* consumption and platform revenues?

To answer this, we analyze the change in aggregate consumption and purchases at the series level over the 15 days before and after the wait-time reduction. We estimate the stacked DiD regression from Equation 4 using a two-period series panel data, where the dependent variables are the total number of episodes consumed and purchased for a given series (log-transformed). We control for a host of observable characteristics that may affect aggregate demand, including series length, days since first and last publication, and the presence of promotions. Again, we include group-series and group-period fixed effects.

The results of our main analysis presented in Table 10 demonstrate a significant positive effect of the wait-time reduction on aggregate consumption and purchases. Column (1) shows the impact on total consumption. The coefficient of *after* $\times$ *treated* suggests that if the episodes consumed before the reduction is at the mean (2610), holding all else equal, it would increase to 5011, a 92% increase. Given our earlier findings, we also expect to find an increase in aggregate purchases. The estimated treatment effect in Column (2) is indeed positive and significant. It suggests that if the purchased episodes before the reduction is at the mean (868), it would increase to 1033, a 19% increase.

The increase in aggregate purchases reflects both the increase in intensive and extensive margins. We decompose the total impact into relative contributions from each dimension in a rough back-of-the-envelope calculation using the estimated parameters. Since purchases before the reduction are from the existing consumers by definition, assume this increases by 0.073 under reduced wait-times, the estimated coefficient of *after* $\times$ *treated* estimated in Table 7. This suggests that aggregate purchases from the existing consumers would increase to 934, representing an 8% increase. The remaining 11% increase is driven by the purchases from the new consumers. In relative terms, the intensive and extensive margins account for 40% and 60% of the increase in aggregate purchases, respectively. This highlights the importance to platforms of encouraging consumer progress through the series to ensure that they fully benefit from episode complementarities.

The results are robust to the wider time window of 60 days and the alternative model specification using magnitude of wait-time reduction. Given the log-log specification, the estimated coefficient for *after* $\times$ *magnitude* yields the demand elasticity with respect to wait-times. We find that a 1% reduction in wait-times lead to 0.26% ($p < 0.01$) and 0.06% ($p < 0.1$) increase in aggregate consumption and purchases, respectively.

	(1) log(Consumed+1)	(2) log(Purchased+1)
after × treated	0.652***	0.174**
	(0.051)	(0.078)
log(no. episodes)	0.888*	0.784
	(0.502)	(0.753)
log(days since first pub.)	-0.166	-0.649
	(0.679)	(0.653)
log(days since last pub.)	-0.029	-0.060
	(0.033)	(0.042)
$\mathbb{1}$(promotion)	0.388***	0.281*
	(0.132)	(0.150)
Group-Series FE	Y	Y
Group-Period FE	Y	Y
N Obs	2292	2292
N Series Groups	191	191
R-squared Adj.	0.185	0.017

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

Table 10: Treatment effect on aggregate consumption and purchases

6.5 Robustness Checks

6.5.1 Identifying Assumptions

Causal identification of the DiD estimate 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 and the model would yield unbiased estimates. In our empirical setting, since the platform reduced the WFF period without notifying users in advance, there was no potential for strategic action from the users such as delaying consumption or purchase.

To formally test the parallel trends assumption, we follow the approach commonly used in the literature that exploits the pre-treatment panel data (Angrist and Krueger, 1999; Bronnenberg et al., 2020). We take the 15-day pre-treatment period and run the analysis from Section 6.4 replacing *after* × *treated* with *period* × *treated*, where *period* indicates day. As shown in Table 11, the deviation from the common trend for the treatment series is very small and not statistically significant. Thus, we fail to reject the null hypothesis that the trend of the treated series is not significantly different from the control series, providing support of the parallel trends assumption. The results are robust to the use of a longer pre-treatment window of 30 days.

We also conduct a sensitivity test following Cinelli and Hazlett (2020) and Gibson and Zimmerman (2021) to assess how much deviation from parallel trends would be required to undermine our conclusions, which are

	(1) log(Consumed+1)	(2) log(Purchased+1)
period × treated	0.003	-0.007
	(0.005)	(0.006)
Other control variables	Y	Y
Fixed Effects	Y	Y
N Obs	17190	17190
N Series Groups	191	191
R-squared Adj.	0.019	0.009

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

Table 11: Parallel trends

based on the assumption of parallel trends. Cinelli and Hazlett (2020) reformulate the classical omitted variables framework to develop a sensitivity analysis that provides, relative to an observed covariate benchmark, how strongly unobserved confounders would need to be associated with both the outcome and treatment variables (in terms of partial R^2) to explain away the estimated treatment effect. The key advantage of using R^2 is that it is scale-free and does not require distributional assumptions of unobserved confounders as well as on the treatment assignment mechanism. As the benchmark observed covariate, we do not rely on a single variable, but rather include all observed covariates from Table 10 to be more conservative. The results presented in Figure 12 show that even if the unobserved confounders are twice as strong as the combined explanatory power of the benchmark covariates for treatment and outcome, the effects remain consistent.

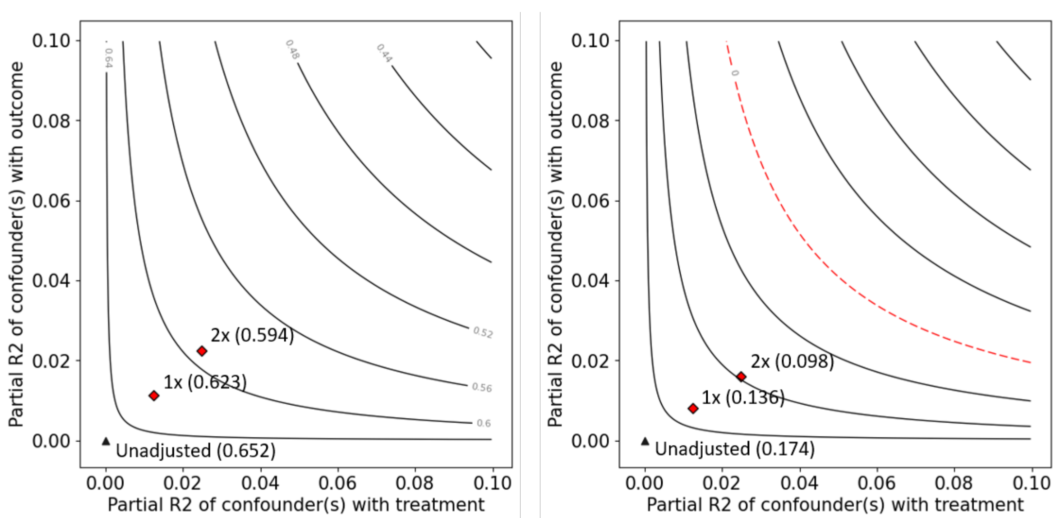

Figure 12: Sensitivity contour plots of treatment effect on aggregate consumption (left) and purchases (right). Black triangle indicates the baseline treatment effect. Red diamonds indicate the treatment effect assuming unobserved confounders 1x and 2x as strong as the combined explanatory power of the benchmark covariates.

SUTVA states that the potential outcomes of each unit are not influenced by the treatment assignment or outcomes of other units. In our context, there may be potential substitution between series where readers move away from non-treated series to spend more time reading series with reduced wait-times. There also may be complementary effects where readers spend more time consuming episodes of the treated series and in doing so also increase consumption of the untreated series. The platform hosts over 10,000 series on the platform, and the treatment was implemented at various times across only 191 series. Hence, it is unlikely that a reduction in wait-time had meaningful spillover effects on other series on the platform.

Nonetheless, we adopt two approaches to mitigate the concerns around the potential violation of SUTVA. Specifically, we show that the results remain robust using (1) a subset of control series with minimal overlap in reader base between the treated series prior to treatment and (2) an alternative model specification that explicitly controls for the overlap. The assumption is that if there are any substitution or complementary spillover effects, they should be greater for the non-treated series who share more readers in common with the treated series. We define $overlap_{sg}$ as the proportion of readers of series s who have also read an episode of the treated series in group g during the 15 days prior to treatment ($overlap_{sg} = 0$ for treated series).

First, we conduct the analysis from Section 6.4 on a subsample dropping all control series with an overlap greater than 10%. Second, we estimate the treatment effects by explicitly controlling for potential spillovers between series, motivated by the approach from Clarke (2017) and Jo et al. (2020). By including an interaction term *after* × *overlap*, we run the following regression which ensures that the treatment effect is isolated from any spillover effects:

$$Y_{sgp} = \beta^{DD}(after_p \times treated_s) + \rho(after_p \times overlap_{sg}) + X_{sgp}\gamma + \delta_{sg} + \nu_{gp} + \epsilon_{sgp} \tag{10}$$

where ρ is the spillover effect from the treated to the untreated series. The treatment effect estimates from both approaches, presented in Table 12 remain qualitatively unchanged. In columns (3) and (4), the estimated coefficient on *after* × *overlap* is not significant, providing empirical support for SUTVA.

	Subsample Analysis		Controlling for Potential Spillovers	
	(1) log(Consumed + 1)	(2) log(Purchased + 1)	(3) log(Consumed + 1)	(4) log(Purchased + 1)
after × treated	0.645***	0.165**	0.638***	0.159*
	(0.051)	(0.078)	(0.053)	(0.083)
after × overlap			-0.363	-0.368
			(0.308)	(0.690)
Other control variables	Y	Y	Y	Y
Fixed Effects	Y	Y	Y	Y
N Obs	2078	2078	2292	2292
N Series Groups	189	189	191	191
R-squared Adj.	0.185	0.016	0.185	0.017

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

Table 12: Check for potential violation of SUTVA using reader base overlap

6.5.2 Falsification Tests

We test the possibility that the estimated treatment effects are coincidentally picking up spurious correlations through falsification tests using pseudo treatment indicators and dates. For pseudo treatment indicators, we randomly assign a control series as treated for each series group and re-estimate the model without the actual treated series. Since the pseudo indicator does not reflect the true information of whether the wait-time of the series is reduced, the estimated treatment effects should not be significant (Ghose and Todri-Adamopoulos, 2016; Jo et al., 2020). For pseudo treatment dates, we adjust the treatment date to be 15 days prior to the actual date of reduction. Since the modified time frame does not include the actual treatment date, the estimates again should not be significant. Table 13 shows that the estimated treatment effects are indeed not significant, indicating that our findings are not a statistical artifact of our specification.

	Pseudo Treatment Indicator		Pseudo Treatment Date	
	(1) log(Consumed + 1)	(2) log(Purchased + 1)	(3) log(Consumed + 1)	(4) log(Purchased + 1)
after × treated	-0.022	-0.023	0.004	0.020
	(0.037)	(0.055)	(0.068)	(0.098)
Other control variables	Y	Y	Y	Y
Fixed Effects	Y	Y	Y	Y
N Obs	1910	1910	2278	2278
N Series Groups	191	191	191	191
R-squared Adj.	0.035	0.013	0.050	0.028

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

Table 13: Falsification tests using pseudo treatment indicators and dates

6.5.3 Alternative Explanations and Model Specifications

Next, we explore whether the estimated treatment effect could arise from the platform strategically timing the wait-time reduction and promotions. If the platform sent coupons in anticipation of or concurrently with the reduction to draw reader attention, then the estimated treatment effect may be confounded by the promotional effect. Although we explicitly control for the presence of promotions throughout our analyses, we re-estimate the model by removing any series (treated or non-treated) that had promotions during the 30-day time frame. Results presented in columns (1) and (2) of Table 14 remain qualitatively unchanged.

We also check for robustness using a subsample of the treated series with the same post-reduction wait-time. As shown in Table 3, there is variation in wait-times before and after the reduction among the treated series. Our main analysis investigates the effect of the reduction irrespective of the starting and ending points, as our goal is to explore the consumption dynamics that arise due to the properties of serialized media rather than pinpoint an optimal wait-time. Nevertheless, there may be heterogeneous effects based on wait-time length that may bias our results. We demonstrate robustness to the model specification using reduction magnitude, but to further alleviate the concern, we select a subsample of 152 treated series for which wait-times were reduced to one hour. Results shown in columns (3) and (4) of Table 14 remain robust.

	Excluding Promotions		1-hr Wait-time Post-treatment	
	(1) log(Consumed + 1)	(2) log(Purchased + 1)	(3) log(Consumed + 1)	(4) log(Purchased + 1)
after × treated	0.629***	0.182**	0.731***	0.202**
	(0.052)	(0.083)	(0.060)	(0.091)
Other control variables	Y	Y	Y	Y
Fixed Effects	Y	Y	Y	Y
N Obs	1970	1970	1824	1824
N Series Groups	176	176	152	152
R-squared Adj.	0.171	0.011	0.215	0.020

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

Table 14: Subsample analysis using series with no promotional activities and 1-hour wait-time post-treatment

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 are all fighting to occupy consumers' free time such that they will be the first in line when they 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 delve into the impact of wait-times on individual consumption decisions and monetization of serialized media. We begin by highlighting the conceptual characteristics unique to a series, namely the directed complementarities between episodes that diminish over the consumption interval and the consumer's varying WTP over time. These characteristics motivate us to hone in on the intensive and extensive margins as well as the consumption pace. Importantly, we hypothesize that allowing quicker free consumption can lead to an increase in the intensive margin, despite the risk of purchase cannibalization.

Using data from a platform serving serialized books, we leverage an exogenous shift in policy where the platform reduced the wait-times for a set of series. We estimate using a difference-in-difference framework how the reduction impacted user consumption and purchases. We provide evidence that the positive across-episode spillovers from complementarity dominate the negative cannibalization effect. In addition, we find evidence of consumption acceleration and the expansion of extensive margins. The net impact is an 92%

and 19% increase in aggregate consumption and purchases, respectively. We conduct a battery of robustness checks to support the identification assumptions and rule out any spurious correlations.

The richer insights into the multifaceted consumption dynamics induced by WFF can aid firms with their monetization strategies. The paper goes beyond the traditional focus on the acquisition-cannibalization trade-off to demonstrate how firms may leverage the complementarities across episodes to boost paid consumption. Our research also underscores the significance of recognizing consumer heterogeneity. For instance, a segment of price-sensitive consumers may not necessarily respond positively to more lenient policies on free consumption. This suggests that platforms could gain more from implementing targeted strategies, tailoring policies to different consumer groups to maximize overall revenues.

We conclude with a discussion of limitations in our analysis. 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 could further strengthen our findings. Second, our analysis demonstrates the causal effect of changing wait-times but cannot comment on the optimal wait-time. This would require a structural analysis. With a rich state-dependent utility model incorporating inherent individual characteristics, consumption context and episode content, it would also be possible to explore optimal policies, such as charging a smaller, but positive price on waited episodes or targeted wait-times. Third, given the importance of episode complementarities, leveraging recent advancements in text analysis to understand how the episode content such as the strength of cliffhangers, level of suspense and sentiment affects consumption decisions may be a fruitful area of future research. Fourth, although we provide ample evidence that correlated marketing activities are an unlikely driver of our results, we cannot fully rule out that firms may have tried to support the treated series with through curation or recommendation algorithms, that could impact 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. Future work should explore the long-term effect of varying wait-times on platform-wide consumption and series that they consume.

Funding and Competing Interests

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

References

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

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

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

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

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

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

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

Cao, J., P. Chintagunta, and S. Li (2023). From free to paid: Monetizing a non-advertising-based app. *Journal of Marketing Research 60*(4), 707–727.

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

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

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

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

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

Clarke, D. G. (1976). Econometric measurement of the duration of advertising effect on sales. *Journal of Marketing Research 13*(4), 345–357.

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

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

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

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

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

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

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

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

Gibson, L. and F. Zimmerman (2021). Measuring the sensitivity of difference-in-difference estimates to the parallel trends assumption. *Research Methods in Medicine & Health Sciences 2*(4), 148–156.

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

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

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

Gu, X., P. Kannan, and L. Ma (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.

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

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

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

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

Kamada, Y. and A. Öry (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.

Klingberg, T. (2009). *The overflowing brain: Information overload and the limits of working memory*. Oxford University Press.

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

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

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

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

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

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

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

Mahajan, V. and E. Muller (1986). Advertising pulsing policies for generating awareness for new products. *Marketing Science 5*(2), 89–106.

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

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

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

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

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

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

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

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

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

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

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