Variety Effects in Mobile Advertising

Omid Rafieian[*] Hema Yoganarasimhan[*]

Cornell University University of Washington

[*] We are grateful to an anonymous firm for providing the data and to the UW-Foster High Performance Computing Lab for providing us with computing resources. We thank Ryan Dew, Simha Mummalaneni, and Hoori Rafieian for detailed comments that have improved the paper. We also thank the participants of the 2018 UW/UBC Conference, 2018 ISMS Marketing Science Conference, 2018 SICS Conference, for their feedback. Please address all correspondence to: or83@cornell.edu, hemay@uw.edu.

Abstract

Users are often exposed to a sequence of short-lived marketing interventions (e.g., ads) within each usage session in mobile apps. This study examines how an increase in the variety of ads shown in a session affects a user's response to the next ad. The authors leverage the quasi-experimental variation in ad assignment in their data and propose an empirical framework that accounts for different types of confounding to isolate the effects of a unit increase in variety. Across a series of models, the authors consistently show that an increase in ad variety in a session results in a higher response rate to the next ad: holding all else fixed, a unit increase in variety of the prior sequence of ads can increase the click-through rate on the next ad by approximately 13%. The authors then explore the underlying mechanism and document empirical evidence for an attention-based account. The paper offers important managerial implications since it identifies a source of interdependence across ad exposures that is often ignored in the design of advertising auctions. Further, the attention-based mechanism suggests that platforms can incorporate real-time attention measures to help advertisers with targeting dynamics.

Keywords: mobile advertising, variety, diversity, attention, digitization, sequential decision-making

Introduction

The smartphone industry has seen unprecedented growth over the past decade, with over three billion worldwide users in 2020 (eMarketer, 2020a). In 2019, an average US adult spent almost four hours per day on mobile devices, surpassing the time spent on TV for the first time (eMarketer, 2019). This growth in smartphone adoption and usage has made mobile an attractive medium for marketing interventions. These interventions are often short-lived in nature (e.g., mobile ads, MMS, push notifications) and provide marketers with many opportunities to interact with users. As such, mobile users are exposed to a variety of marketing interventions, even within a short period of time.

As the variety of marketing interventions increases, it is important for managers and researchers to understand the consequences of this increased variety in the mobile ecosystem. So far, the literature has viewed the increased variety of interventions as a means to explore (vs. exploit) and learn consumers' tastes (Lattimore and Szepesvári, 2020), increase fairness (Dwork et al., 2012), prevent polarization (Celis et al., 2019), and increase reachability (Dean et al., 2020). However, it is not clear how the increased variety itself affects consumer behavior – what are the behavioral consequences of increasing variety that are reflected in consumers' choice? The answer to this question is particularly relevant for managers and researchers as it adds to our understanding of the consequences of this increased variety in the mobile ecosystem, and has important market design implications for the platforms that design marketplaces for consumers and marketers to interact.

In this paper, we study the effects of increased ad variety on consumer behavior in the context of mobile in-app advertising, the most popular type of marketing intervention on mobile devices. It is now the dominant advertising channel in the United States and generates over $100 billion in ad spend (eMarketer, 2020b). A common feature of in-app advertising is the use of refreshable ad slots: each ad exposure lasts a short amount of time (e.g., one minute), and the slot is then refreshed with another ad. Hence, even in a short ten-minute session, a user can be exposed to a large variety of different ads. This feature of in-app advertising makes it particularly suitable for studying how the variety of the sequence of ads seen earlier in the session affects a user's responsiveness to the current ad. Specifically, we seek answer the following questions:

1. How does an increase in ad variety in a session influence users' responsiveness to the next ad?
2. What is the underlying mechanism that explains these effects (if any)?
3. What are the managerial implications of these findings for platforms?

To answer these questions, we use large-scale data from the leading mobile in-app ad-network from a large Asian country. Two key features of the ad-network make it well-suited to study variety effects. First, like most in-app ad networks, it employs a refreshable ad format, where each ad lasts one minute and is followed by another ad exposure. Second, it uses a probabilistic auction to allocate ads, which allows for a wide variety of ads to be shown in a single session. More importantly, it gives us unconfoundedness in ad assignment, a condition that is necessary for causal inference (and is often missing in observational studies on advertising). Together, these two features provide the necessary requirements to study the main question in the paper from a platform's perspective – how does a unit increase in ad variety in a session influence users' responsiveness to the next ad?

We face three key challenges in satisfactorily answering our research questions. First, our treatment variable – an increase in the ad variety of a session – is not fully randomized. Some impressions have a higher propensity to be assigned to the treatment than others due to pre-treatment characteristics such as targeting variables and the set of prior ads. We refer to this as the pre-treatment confounding problem. Second, we are interested in the outcome at the next ad exposure after the treatment. However, some users may leave the session right after receiving the treatment, before we observe the outcome. In particular, if their decision to leave is a function of their treatment, this censoring can interfere with inference. We refer to this as the dynamic selection problem. Third, during the post-treatment phase – from the treatment assignment to outcome collection phase – other variables can also co-vary with our treatment (e.g., the identity of the ads shown, their recency and frequency), thereby making it challenging to isolate the treatment effect of an increase in ad variety. We call this issue post-treatment confounding, since our treatment definition is not separate from some other post-treatment factors that might also affect users' decisions.

We present a methodological framework that uses the uncounfoundedness of ad assignment and helps us address the aforementioned challenges. First, to account for the non-randomness in assignment to an increase in ad variety (pre-treatment confounding), we show that the unconfoundedness of ad assignment implies the unconfoundedness of our treatment. That is, conditional on observables, treatment assignment is as good as exogenous. Specifically, we use the main insight from Rosenbaum and Rubin (1983) and estimate the propensity score for being assigned to an increase in ad variety for any given impression. We then assess covariate balance for these propensity scores and feed them into the main regression model to control for pre-treatment confounding.

Second, to address our dynamic selection challenge, we employ an imputation strategy where we impute the observation in the next period for users who have left the session after being assigned to treatment (Little and Yau, 1996). We show that under the unconfoundedness of ad assignment, we can accurately impute the ad that would have been shown had the user not left the session, because we can estimate the distribution of ad assignment given observables. In particular, we use a specific feature of our setting that the auction would be identical for two impressions that happen around the same time and share the same targeting characteristics. This allows us to use a complementary sample of auctions that happened around the same time as our missing impressions and impute the missing ad assignments.

Our solution to the third challenge of post-treatment confounding is based on a simple logic: we want to ensure that our estimates of treatment effect only capture our treatment – an increase in the ad variety of the session. As such, we need to control for any post-treatment information that imperfectly co-varies with our treatment. Since our post-treatment phase spans over two consecutive exposures (often referred to as exposure $t-1$ and t) where the treatment is assigned in the former, and the outcome is collected in the latter exposure, we control for the fixed effects of the specific ads shown in these two exposures, as well as the frequency and recency of the ad shown in the later exposure. This allows us to achieve a *ceteris paribus* interpretation under some mild conditions. That is, if our treatment effect estimate is β, the interpretation would be as follows: a unit increase in the ad variety of a session increases the outcome on the next ad by β, holding all else constant.

We use an inverse probability weight-adjusted regression to estimate our main effects. We find that an exogenous increase in the ad variety in a session results in a significantly higher click-through rate (CTR) on the next ad, holding all else fixed. The magnitude of our treatment effect accounts for approximately 13$\%$ of the average CTR, which implies that variety effects are relatively sizable. We then consider a series of alternative specifications to examine the robustness of our findings. Notably, we consider a restrictive exact matching model, where we match the exact sequence of ads shown in a session except the ad shown in the treatment stage (e.g., matching $\langle A, B, A, C, D\rangle$ from treatment to $\langle A, B, A, B, D\rangle$ from the control) and control for the fixed effects of the ad that is different so we can fully isolate the effect of an increase in ad variety. Across all the robustness checks, our results show the same pattern – increasing ad variety generates more clicks on the next exposure.

Next, we explore the mechanism underlying our main findings. In our analysis, we focus on the main feature that differs across our treatment and control groups – the novelty of the ad shown at the treatment phase. As echoed extensively in the behavioral literature, showing more novel stimuli in a given space increases users' attention to that space (Helson, 1948; Kahneman, 1973). Therefore, we develop an attention-based explanation for variety effects, wherein the novelty of prior ads increases users' attention to the advertising slot, which, in turn, increases their likelihood of clicking on the next ad.

We conduct a series of empirical tests to examine the validity of the predictions delivered by this mechanism. First, we use the fact that the novelty of the control condition (a non-increase in ad variety) can vary based on the within-session frequency and recency of the ad shown in the treatment phase: the more frequent and recent the ad is, the less novel we expect it to be. We show that as the novelty of the control group drops, the treatment effects become larger. Second, users with higher levels of pre-session exposure to ads and those with more recent pre-session ad exposures exhibit smaller treatment effects. This is consistent with the idea that the novelty of within-session interventions deteriorates with higher and more recent ad exposures prior to the session. Third, if our attention-based account is correct, the treatment effect should be smaller when the past variety is already high because the novelty of an increase in variety is less likely to have an impact on user attention in this case. We test this prediction and demonstrate that the treatment effects are indeed smaller when the past variety is already high and vice versa. Together, these tests provide empirical support for the validity of the proposed mechanism.

In sum, our paper contributes to the literature in several ways. Substantively, the main contribution of the paper is in establishing the causal effects of variety in the advertising context. To our knowledge, this is the first paper to study the downstream effects of an increase in ad variety on consumer's response.[1] This broadens our understanding of how constructs such as variety and diversity affect consumer behavior, which is of critical importance to digital platforms as they employ more experimentation techniques and commit to increasing ad diversity. Further, we propose an attention-based mechanism to explain the effects of an increase in ad variety. This finding is important as it provides a parsimonious and testable account on the effects of variety, and highlights the importance of attention-based measures. One key implication for platforms is to incorporate these attention-based measures in their targeting offerings to advertisers as well as their quality scoring system to improve the performance of their auctions. From a methodological standpoint, our paper develops an empirical framework to study the effects of variety in sequential settings that can be applied to other domains (e.g., music streaming services). Our empirical framework is fairly general as it only requires the unconfoundedness assumption that can be easily satisfied by a digital platform. Overall, we expect our substantive findings and empirical framework to be of relevance to

[1] The only prior study related to variety in this context is Schumann et al. (1990), which shows the variation of ad content over a repeated advertising schedule will increase user's responsiveness to that ad. While they only focus on variation in the content for one ad in a lab setting, our work extends it to the variety of potentially competing ads in a large-scale in-app advertising market.

digital platforms within and outside the advertising domain.

Related Literature

First, our paper relates to the marketing literature on variety. The concept of variety has been examined through two broad viewpoints. The first stream views variety (in consumption) as an outcome variable and studies consumers' variety-seeking behavior (McAlister, 1982; Ratner et al., 1999), and more broadly their demand for variety in products (Kim et al., 2002; Datta et al., 2017). In the second stream, variety serves as a factor influencing some outcome variables associated with consumer behavior, such as the link between the variety of assortments and store choice (Hoch et al., 1999), variety of episodes and consumer's engagement (Redden, 2008), and the dispersion of word-of-mouth and TV ratings (Godes and Mayzlin, 2004). In line with the second stream, our paper studies the effects of increasing ad variety on consumers' ad response. Our work contributes to this literature in two ways. From a substantive viewpoint, we add to this literature by establishing variety effects in the context of advertising, and proposing an attention-based account to explain these effects that can be applied to other domains. From a methodological point-of-view, we extend this literature by providing an empirical framework to study the effects of variety in sequential treatment settings.

Second, our paper relates to the literature on advertising marketplaces that adopts a platform perspective to study the questions related to advertising. The work in this domain often focuses on broad questions of market design (Yao and Mela, 2011; Choi and Mela, 2019), platforms' incentives to provide tools such as granular behavioral targeting (Rafieian and Yoganarasimhan, 2021) or ad avoidance technology (Wilbur, 2008; Tuchman et al., 2018), and advertising externalities that affect market design (Wilbur et al., 2013). Our paper extends this literature by recognizing a new type of externality created by ad variety, which is of relevance to advertising marketplaces that seek to incorporate diversity and fairness criteria in their decision-making. Besides, our paper highlights important challenges in designing auctions and optimal adaptive experimentation systems in light of this externality and offers potential solutions.

Finally, our paper relates to the literature on advertising dynamics. Prior literature on TV advertising has extensively documented evidence for different dynamic effects such as ad avoidance, carryover effects, wear-in, wear-out, and S-shaped ad response curve (Danaher, 1995; Naik et al., 1998; Tellis, 2003; **?**; Aravindakshan et al., 2012). While this body of work focuses on aggregate response models, a series of recent papers on digital advertising take advantage of individual-level data and document temporal effects such as carryover effects and wearout (Chae et al., 2019), effect of multiple ad creatives (Braun and Moe, 2013; Bruce et al., 2017), advertising avoidance (Wilbur, 2016; Deng and Mela, 2018), spillover effects (Rutz and Bucklin, 2011; Jeziorski and Segal, 2015; Sahni, 2016), and effects of temporal spacing (Sahni, 2015), and more broadly attribution dynamics (Li and Kannan, 2014; Zantedeschi et al., 2017; Danaher and van Heerde, 2018). Please see Bucklin and Hoban (2017) for an excellent summary of individual-level models of digital advertising. Our paper extends this stream of literature in two ways. First, we establish a new dynamic effect, i.e., the effect of an increase in ad variety. Second, we offer a new behavioral account of consumers' ad response based on attention and adaptation-level theory that can be used more broadly in the studies of advertising. Finally, we provide a methodological framework that can be used in other studies on dynamics of advertising under some mild assumptions.

Setting and Data

We now describe our setting, data and sampling, data generating process, and present some summary statistics and descriptive analysis.

Setting

Our data come from the leading mobile in-app advertising network of a large Asian country, which had over 85% market share during the time of this study. The ad-network functions as a match-maker between advertisers and mobile apps and serves ads inside mobile apps. The scope and scale of the ad network are quite large, and it generates a total of over 50 million ad impressions daily. We begin by describing the four main players in this marketplace:

- *Users:* Individuals who generate eyeballs or impressions by using their mobile apps. For each

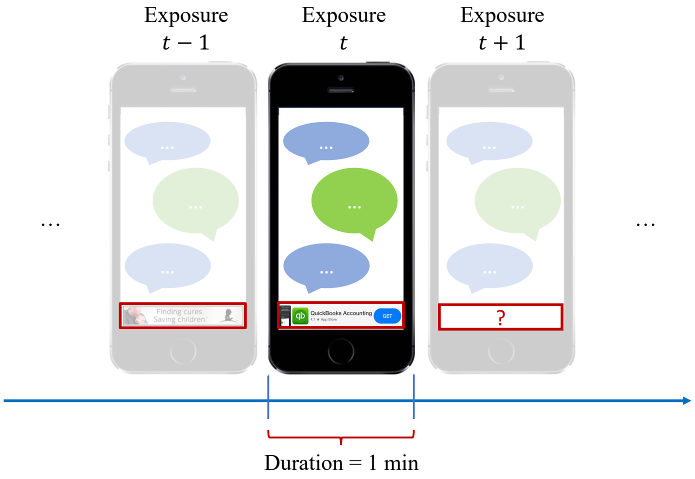


Figure 1: A visual representation of the ad slot in our setting. The highlighted rectangle at the bottom of app is the refreshable ad slot.

ad impression shown to her, the user can decide whether to click on the ad or not.

- *Publishers:* App developers who have joined the ad-network. Their revenue comes from the clicks generated in the ads shown in their apps. This is the main monetization strategy for most of the apps in our data.
- *Advertisers:* Firms (usually mobile apps) who wants to show their ads to mobile app consumers. Advertisers create banner ads (in either JPEG or GIF formats), submit a bid that indicates their willingness to pay per click, and a maximum daily budget. Advertisers can also target their ads based on the following variables associated with an impression: province, app category, hour of the day, smartphone brand, mobile service provider, and connectivity type.
- *Ad-network:* The platform that facilitates the matching between ads and impressions generated by a user-app combination. It runs a real-time auction to select the ad to show in each impression. In our context, the platform has full control over ad format (only banner ads in the bottom of apps' screen) and allocation through the auction. It operates based on cost-per-click (CPC) revenue model, which indicates that it generates revenue only when a click happens. As such, like most platforms, one of their key economic incentives is to use the understanding of the drivers of clicks to generate more revenue.

Our ad-network employs refreshable ad-slots, where each impression lasts one minute.[2] When a user starts using an app, the ad-network runs an auction to determine the winning ad and serves this ad for one minute. If the user continues using the app beyond one minute, the ad-network treats this as a new impression and runs another auction to determine the next ad to show the user. The practice of using a refreshable ad slot helps create more ad impression opportunities by reducing the time allocated to each impression. Figure 1 presents a visual representation of the ad format, ad slot, and the sequential nature of ad delivery in our setting.

Data

We have data on all the impressions and corresponding clicks (if any) in the platform for all participating apps for a one-month period from 30 September 2015 to 30 October 2015. For each impression, we observe the following information: (1) Time and date, (2) AAID (user identifier), (3) app ID (publisher), (4) ad ID, (5) bid submitted by the winning ad, (5) GPS information including the exact latitude and longitude of the user, (6) click indicator, and (7) targeting variables that contain the province, app category, hour of the day, smartphone brand, connectivity type, and mobile service provider (MSP). Notably, all the variables that the advertiser can target are observable to us.[3] Thus, we are able to avoid many of the common problems related to endogeneity in the measurement of ad effects due to targeting based on unobservables, as well will discuss later.

Overall, the scale of our data is quite large, with over 1,594,831,699 impressions over a one month period with 600 auctions per second on average. We now explain how we sample from this data and then describe an important aspect of the data-generating process – the auction mechanism.

Session Definition and Sampling

Recall that our goal is to study the effects of an increase in ad variety in a session on the user's response to the current ad. As such, the definition of a session is central to our study. Since we

[2] It is important for the length of all the exposures to be the same, since we know that length of ad exposures affects ad responsiveness in digital advertising settings (Danaher and Mullarkey, 2003).

[3] The ad-network also has access to the ISP for each impression if it happens on WiFi. However, in our data, this information is missing for the vast majority impressions and advertisers do not seem to use it for targeting. Hence, we do not use ISP in our analysis.

focus on an increase in ad variety in a session, we want the user in a session to be able to recognize this increase. Hence, we need to avoid long usage gaps in our session definition to ensure that the user would be able to recognize an increase in ad variety. To that end, we define a session as a set of consecutive impressions generated by a user within an app, such that the gap between two consecutive impressions is less than 10 minutes.[4]

Given the substantive goal of our research, we construct a sample of users for whom we have we have the entire behavioral history on the platform. This excludes the sample of users whose past activity goes before the beginning of our data, as well as those whose activity logs are not stored in the platform server at different points. We then focus on the top app and collect the data of all sessions with our sample of users using this app.[5] Overall, it gives us a sample of 85,450 users who generate 1,197,850 sessions and 6,805,322 impressions in the messenger app. We see a total of 327 unique ads in this sample. The length of the sessions in our sample varies quite a bit. While half of the sessions end after the first two exposures, over 25% last five or more exposures (see Figure A1 in Web Appendix A).

All the descriptives in this section are shown for this sample. However, we utilize the data from other users and apps to supplement our analysis. For the sample of users that we focus on, we keep track of the data generated by these users in other apps to segment them based on their behavioral history and explore the heterogeneity in their responsiveness to ad variety. Further, we also use the impressions from other users not in our sample in the top app for imputation purposes in our estimation procedure. We discuss these uses in detail later.

Finally, it is worth emphasizing that although we use the same data source as Rafieian and

[4] We do not assume that a click automatically ends a session because of two reasons. First, empirically, we see that 80% of users who click come back to the app within 10 minutes. Second, there is no theoretical reason to believe that a click would affect user's memory of prior ads if they came back to the app within a short time. That said, our results are robust to alternative definitions of a session, e.g., when we assume a click ends a session, or when we allow for larger or smaller gaps between consecutive impressions.

[5] Our choice of using the top app is only for cleaner analysis, where we can better control the context and app-switching. This allows us to only sample sessions that are entirely in one app. The main results of the paper are robust when we consider other apps.

Yoganarasimhan (2021), the specific sample and the goal of the studies differ a lot. Rafieian and Yoganarasimhan (2021) use a random sample of users over the span of October 28–30 to predict CTRs across all apps and ads. They use the earlier data to construct detailed behavioral and contextual features and then examine the platform's incentives to allow more granular targeting in counterfactual scenarios. Thus, they focus on the interplay between the platform's market design and advertisers' bidding decisions in a two-sided market. On the other hand, we have an entirely different goal in this paper. We want to examine a consumer-level effect when the platform increases ad variety. Further, we use a completely different sample where we can make sure we have the entire behavioral history for all of our users and focus our analysis on the data from the top app.

Data Generating Process

We now describe the data generating process in our setting. Let i denote a session, and t denote an exposure number within a session. Each exposure t in session i comes with three pieces of information: (1) impression-level characteristics ($X_{i,t}$) that capture all the observable attributes associated with the user and context of the impression (e.g., the smartphone brand), (2) the ad shown in the exposure ($A_{i,t}$), and (3) the user's decision to click on the ad shown in the exposure ($Y_{i,t}$).

The ad-network uses a *quasi-proportional* auction to allocate ads to impressions (Mirrokni et al., 2010). The main distinction between a quasi-proportional auction and other commonly used auctions (e.g., second price or Vickrey) is the use of a probabilistic winning rule, i.e., all ads participating in the auction for an impression have a non-zero probability of winning. For exposure t in session i, the probability that ad a wins this impression is given by:

$$\pi_{i,t}(a) = \mathbb{1}(a \in \mathcal{C}_{i,t}) \frac{b_{i,a} m_{i,a}}{\sum_{k \in \mathcal{C}_{i,t}} b_{i,k} m_{i,k}}, \tag{1}$$

where $\mathcal{C}_{i,t}$ is the set of ads participating in the auction for this exposure, and $b_{i,a}$ and $m_{i,a}$ are advertiser a's bid and quality score in session i, respectively. Thus, the variation in $\pi_{i,t}(a)$ can stem from variation in $\mathcal{C}_{i,t}$, $b_{i,a}$, $m_{i,a}$, or some combination of these variables. The variation in $\mathcal{C}_{i,t}$

is largely driven by advertisers' targeting decisions. For example, if ad a chose not to target the province where the user generating session i is located, then ad a will not belong to $\mathcal{C}_{i,t}$, which implies $\pi_{i,t}(a) = 0$. The quality score $m_{i,a}$ is a measure of profitability that the platform assigns to ad a in session i. The extent of customization in the quality scores is quite low: the ad-network simply assigns one aggregate quality score to each ad and only updates it once a day. So while $m_{i,a}$ can vary across sessions, the extent of variation is quite low. Finally, $b_{i,a}$ is ad a's bid for session i. In our setting, each ad could submit only one bid at a given time of the day.[6]

In sum, $\mathcal{C}_{i,t}$, $b_{i,a}$, and $m_{i,a}$ together determine the distribution of propensity scores ($\pi_{i,t}(a)$s) for ads for each exposure in session i. As such, the ad shown at exposure t, $A_{i,t}$, is a draw from this probability distribution. This probabilistic allocation rule thus gives us random variation in ad assignment across and within sessions, which will form a core part of our identification strategy. It is worth emphasizing that the extent of exogenous variation created by the auction only helps us without limiting our ability to extend our results to other more commonly used auction settings.[7]

Summary Statistics

We now present summary statistics for the key variables of interest and some descriptive analysis.

Targeting Variables

Targeting variables are the dimensions on which advertisers can target their ads. In our setting, these consist of province, app category, hour of the day, smartphone Brand, MSP, and connectivity type. All these variables are categorical, and advertisers can specify which subcategories they want to show their ads in, e.g., an advertiser can indicate that she wants her ads to be shown only from 6 pm to 9 pm every day on Samsung phones in one specific province. We first report the impression share of the top three subcategories within each targeting variable in Table 1.

[6] Advertisers could not customize bids by targeting variables. For example, they could not submit different bids for two different provinces at the same time, even if their willingness to pay for the clicks in the two provinces was different. Further, if an advertiser changes her bid at some point in time, it is updated for all the sessions that start in the next hour of the day.

[7] It is similar to the case where we want to generalize the results from an experiment in Facebook to the setting with their auctions in place. Please see Tunuguntla and Hoban (2021) for a description of commonly used auction mechanisms in digital advertising.

Variable	Number of subcategories	Share of top subcategories			Total number of impressions
		1^{st}	2^{nd}	3^{rd}	
Province	31	24.67%	9.61%	7.45%	6,805,322
Hour of the day	24	8.43%	7.98%	7.21%	6,805,322
Smartphone brand	7	44.62%	38.18%	8.62%	6,177,053
Connectivity type	2	50.33%	49.67%		6,805,322
MSP	3	50.18%	43.98%	5.84%	6,635,836

Table 1: Summary statistics of the targeting variables. The last column shows the total number of non-missing observations for each variable. While the information about province and hour of the day is always available, other variables are missing for some impressions. The shares shown are computed after excluding the missing observations for each variable.

Next, we examine the extent of targeting that occurs in the platform. We know that if a subcategory (e.g., Samsung in the category of smartphone brand) is excluded by an ad through its targeting decision in a given hour of the day, then zero impressions of that ad will be shown in that subcategory during that period. So the number of distinct ads shown within a subcategory is informative of the number of advertisers targeting that subcategory. Further, we can correlate the impression-share of a targeting subcategory with the number of distinct ads targeting it to illustrate the relationship between a subcategory's popularity and the extent to which it is targeted. Therefore, for each subcategory within a targeting variable/dimension, we first calculate the number of distinct ads that show at least one impression in that subcategory. Then, we plot the number of distinct ads targeting it against its share of impressions in that subcategory. The results of this analysis are shown in Figure 2.

Three important patterns emerge from Figure 2. First, we find that some variables are not used much for targeting. In particular, all the subcategories within connectivity type and MSP are close to the grey dashed line at the top. This implies that most advertisers were showing their ads irrespective of these variables. While the subcategories in the smartphone brand differ slightly in the number of ads targeting them, the extent of targeting is still limited. On the other hand, province and hour of the day are the main variables used for targeting: all subcategories within these variables are considerably different in terms of the number of distinct ads targeting them. The second insight from Figure 2 also relates to this difference: subcategories with a higher share of

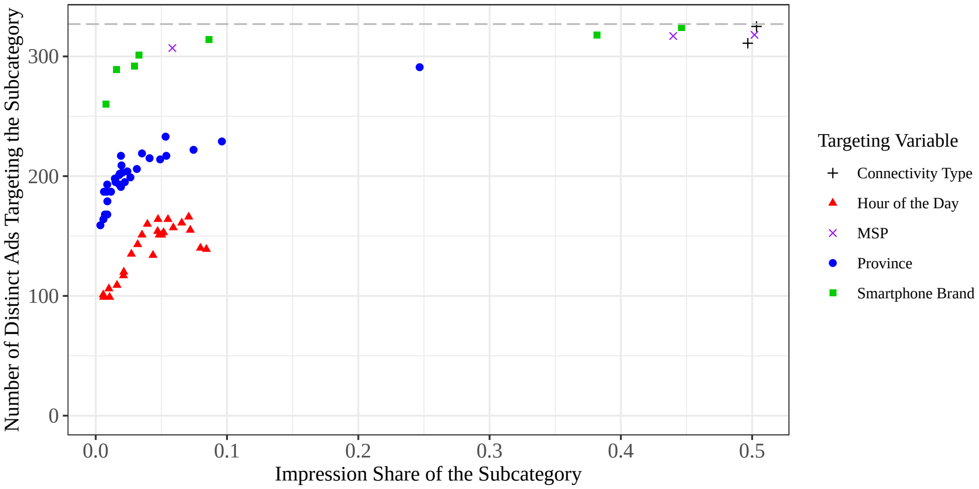
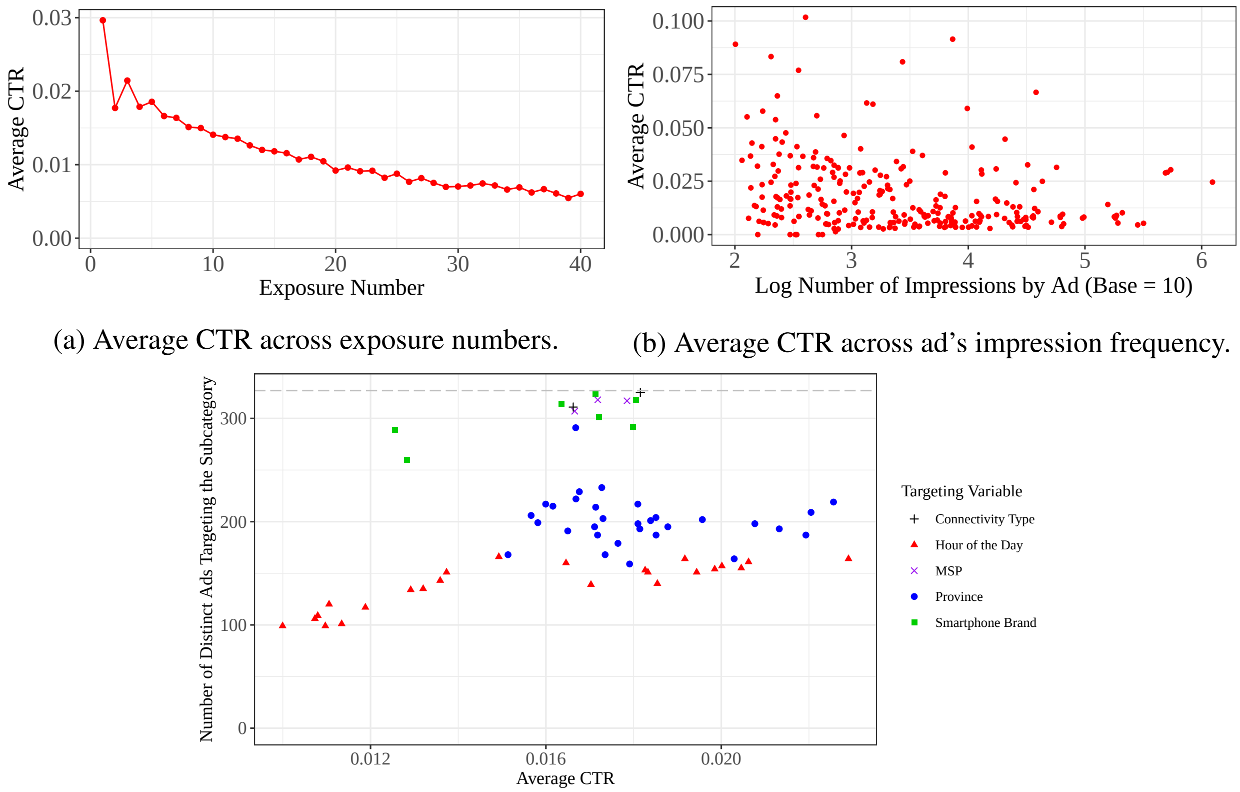


Figure 2: Relationship between the number of distinct ads targeting a subcategory and the impression share of that subcategory within the targeting category. All subcategories within each targeting category are in the same color and shape. The dashed grey line on the top is the total number of distinct ads available in our data (which is 327).

impression within a category seem to be more popular among advertisers. For example, the blue circle in the top-center denotes the largest province in the country (and contains the capital city) with the highest share of impressions. We can see that this province also has the highest number of distinct ads targeting it (compared to other provinces). In contrast, the red triangles in the bottom left are midnight hours that have the lowest impression shares and fewer advertisers targeting them. However, it is worth noting that even these unpopular hours attract a lot of ads (over 100). This brings us to the third key insight from Figure 2: there is no niche targeting in this market. Therefore, we can expect a significant amount of within-session variation in the set of ads in most sessions. This, in turn, facilitates the study of variety effects on user behavior.

Variation in Click-through Rate (CTR)

Click is the main outcome of interest in this study and are are important in our setting for a few key reasons. First, we view the problem through the lens of a platform that runs a CPC auction and wants to generate more clicks. Thus clicks are directly tied to platform revenues. Second, the vast majority of ads in our setting are mobile apps interested in app installs. These ads are often referred to as "performance ads", since they have objective performance measures, as opposed to "brand



(a) Average CTR across exposure numbers.

(b) Average CTR across ad's impression frequency.

(c) Variation in average CTR across targeting subcategories.

Figure 3: Variation in CTR (measure of CTR is in absolute terms not percentage).

ads" whose goal is often to generate more reach and brand recognition (Arnosti et al., 2016). While a click on a brand ad is often not informative of the final conversion outcome (e.g., click on a Ford ad takes the user to Ford's main homepage), a click on a performance ad is a direct step towards conversion. For example, in our context, a click on the advertised app takes the user to the app store page, one click away from conversion. Thus, a click serves as a strong engagement signal of the ultimate outcome for ads in our study.

Figure 3 presents some descriptives on the variation in CTR across: (1) exposure numbers, (2) ads, and (3) sessions. We start with the most basic graph that illustrates the within-session variation in CTR – variation across exposure numbers. Figure 3a reveals a downward trend in average CTR against exposure numbers. This suggests it is essential to control for the exposure number when studying the effects of within-session interventions such as variety.[8] Next, we focus on the variation

[8] The average CTR across all the impressions (for all the exposure numbers) is 0.0174.

in average CTR across ads. In Figure 3b, we plot each ad's average CTR across the log of the number of impressions of the ad. This allows us to visualize the dispersion in ad-specific CTR at different levels of impression frequency. Overall, we observe considerable variation in ad-specific CTR at all frequency levels, especially across low-frequency ads. Finally, we examine the variation in CTR across sessions with respect to their targeting variables. As such, Figure 3c plots the number of distinct ads targeting each subcategory against their average CTR. This is the equivalent of Figure 2 with the difference that the x-axis is the average CTR of each subcategory as opposed to impression share. First, by looking only at the x-axis, we find quite a bit of variation in CTR across provinces and hours of the day. However, the variation in CTR is minimal in the subcategories of other targeting variables such as MSP and connectivity type. In general, the variation in CTR shrinks as the values in the y-axis increase, where almost all ads target all subcategories. Intuitively, this makes sense because advertisers want to use variables that have considerable within-variation in CTR for targeting. Finally, we observe a slightly positive correlation between the number of distinct ads targeting a subcategory and the corresponding CTR, which suggests that subcategories with higher CTRs are more popular among advertisers. This pattern is indicative of a selection problem, which we discuss in greater detail later.

Preliminary Analysis

In this section, we present some preliminary analysis on how users' responsiveness to the current ad relates to the variety of previous ads seen by the user in a session. We first define a simple measure of variety and show some summary statistics on the extent and sources of variation in this measure of variety. We then present some preliminary regression results to demonstrate the patterns in the data on the link between variety of previous ads and user response to the current ad. Finally, we discuss the limitations of our preliminary analysis and why we need a more advanced analysis to establish a causal link between variety and user response.

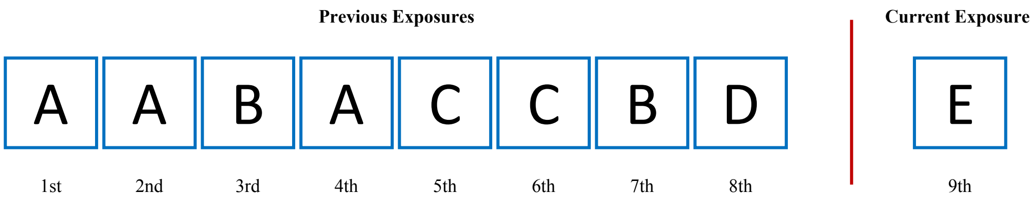


Figure 4: An example of a session where the user is at the ninth exposure. The numbers represent exposure number t, and each rectangle represents the ad shown in that exposure. The letter coding refers to ad IDs, i.e., each letter represents one unique ad. For instance, the user is shown the same ad (coded in letter A) during the first, second, and fourth exposures.

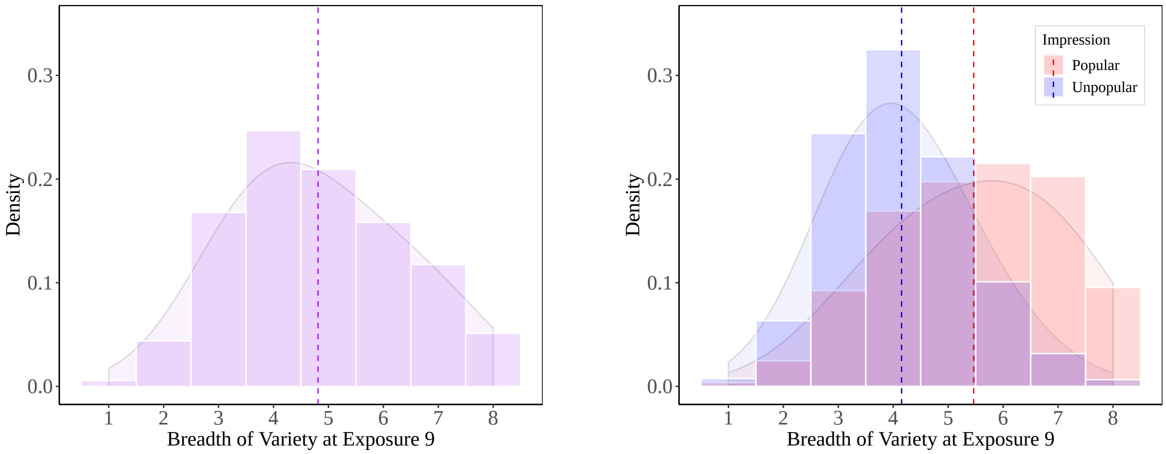
A Simple Measure of Variety

Quantifying variety is a challenging task because consumers' perceptions of variety can vary depending on the context and the information structure of the assortment (Hoch et al., 1999). Different measures in the literature capture certain aspects of the variety in a set of objects, such as the breadth of variety, diversity, or concentration. In our setting, we want to measure variety for the sequence of prior ads shown in a session. Figure 4 presents an example of such a sequence, where a user has seen a sequence of eight ads and is now at the ninth exposure.

In this section, we focus on the simplest conceptualization of variety over a sequence of ads – breadth of variety, which basically counts the number of distinct ads shown. In the example shown in Figure 4, the breadth of variety is four as there are four different ads shown in the sequence of eight prior exposures. We define the sequence of ads shown in session i as $\langle A_{i,t} \rangle_{t=1}^{T_i}$, where $A_{i,t}$ is the ad shown in the t^{th} exposure in session i, and T_i is the total number of exposures shown in session i. We can define the breadth of variety as follows:

$$V_{i,t} = |\{A_{i,1}, \ldots, A_{i,t-1}\}|. \tag{2}$$

Besides simplicity, another advantage of this measure is that we can decompose it into binary pieces. We later use this feature of $V_{i,t}$ to define our empirical problem.



(a) All impressions.

(b) Popular vs. unpopular impressions.

Figure 5: Distribution of breadth of variety at the ninth exposure for all impressions and subsets of popular and unpopular impressions. Popularity is defined as the number of ads that targeted an impression based on the targeting characteristics available.

Variation in Breadth of Variety

We now present some descriptive statistics on the distribution of breadth of variety in our data. Conceptually, the variation in breadth of variety in our setting stems from the probabilistic allocation mechanism. In Figure 5a, we present the empirical distribution of breadth of variety for the ninth impression in the sessions in our data. We see that there is considerable variation in users' exposure to variety: all levels of variety – from one to eight – occur in the data. This is promising since we need sufficient variation in variety to be able to conduct any meaningful analysis.

While we have sufficient variation in users' exposure to variety at any given exposure[9], it is not clear whether this variation is distributed identically across different sessions. Indeed, we know that targeting subcategories vary significantly in their popularity based on advertisers' targeting decisions (as shown in Figures 2 and 3c). As such, we expect sessions with more appealing targeting characteristics to have a higher variety of ads simply because they have more ads in their inventory. We now use this intuition to examine the differences in the distribution of variety across sessions as a function of their targeting popularity. We define the targeting popularity of an impression as the number of ads that are targeting all the subcategories associated with that impression. We then use a

[9] We only show the histogram for the ninth exposure, but the coverage is the same for other exposures.

median split to divide impressions at the ninth exposure into two subsets – popular and unpopular impressions. The distribution of variety for each subset is shown in Figure 5b. This figure confirms our intuition that the two distributions are largely different: more popular impressions show a higher variety of ads with a gap of over one point in the means of these distributions. Thus, in our main analysis, we need to ensure that we separate the effects of variety from targeting popularity.

Preliminary Results

We now run a series of regressions to explore the relationship between the click outcome and the variety of previous ads. Let $Y_{i,t}$ denote the click outcome for session i at exposure t. Further, let $X_{i,t}$ denote the set of pre-session variables (e.g., province, hour of the day)[10], and let $H_{i,t}$ be the sequence of all ads shown in the session (i.e., $H_{i,t} = \langle A_{i,1}, \dots, A_{i,t} \rangle$). For our preliminary analysis, we estimate the following regression model:

$$Y_{i,t} = \beta V_{i,t} + f(X_{i,t}, H_{i,t}) + \epsilon_{i,t}, \tag{3}$$

where β captures the marginal effect of the breadth of variety of previous ads on the user's click probability, and $f(X_{i,t}, H_{i,t})$ can be any non-parametric function that separates out the effects of the other covariates on the outcome from the effects of variety. In our preliminary analysis, we consider different parametric specifications of the function f to estimate the coefficient of variety.

We present our preliminary results in Table 2. We focus on the sample of impressions from the fourth to tenth exposures in a session. The motivation behind our choice of the fourth exposure as the starting point is simply to have more variation in variety. On the other hand, we use the tenth exposure as the ending point only to ensure that we have enough observations per exposure number. In the first column, we consider the most basic model, where we simply regress the outcome on variety, controlling for the ad and exposure number fixed effects. As shown in the first column of Table 2, the coefficient of variety is positive and statistically significant. In light of Figure 5b, we

[10] The reason why we have subscript t for pre-session variables is that some variables such as hour of the day can actually change within the session.

	Dependent variable: Click ($Y_{i,t}$)			
	(1)	(2)	(3)	(4)
Variety ($V_{i,t}$)	0.00245***	0.00107***	0.00107***	0.00084***
	(27.62)	(11.27)	(11.24)	(8.18)
$Freq_{i,t}$				-0.00041^{**}
				(-3.11)
$Space_{i,t}$				0.00031***
				(4.78)
Ad FE	✓	✓	✓	✓
Exposure Number FE	✓	✓	✓	✓
Targeting Variables FE		✓	✓	✓
Session Length FE			✓	✓
R^2	0.007	0.009	0.009	0.009
Adjusted R^2	0.007	0.008	0.009	0.009
No. of Obs.	1,993,542	1,993,542	1,993,542	1,993,542
Note:	*p<0.05; **p<0.01; ***p<0.001			

Table 2: Preliminary results on the effect of breadth of variety on click outcome.

know that the assignment to variety is confounded by advertisers' targeting. Therefore, in the model in the second column, we control for all the targeting variables presented in Table 1. While the magnitude of the variety coefficient changes, the sign and significance remain unchanged.

Next, we focus on another potential confound – session length. Given the sequential nature of the variety assignment, some users may drop out in the middle of the session, i.e., the session length is not the same across all users. In particular, if their decision to leave is influenced by the variety assignment, this would create a dynamic selection issue. A simple (yet insufficient) solution would be to estimate the effects of variety holding the session length constant. This makes sure that we only compare users who made the decision to leave the session at the same point. In the third column of Table 2, we control for the session length in addition to prior controls. The estimate shows the same pattern – the coefficient of variety remains positive and significant.

Finally, note that changes in variety within a session do not happen in isolation. That is, other characteristics of the session (that influence the click outcome) can co-vary with variety of previous ads. As such, without proper controls, the coefficient of variety may actually pick up the effects of these session-level variables rather than variety. Therefore, in the fourth column of Table 2, we add two session-level controls: (1) $\mathit{Freq}_{i,t}$ – the number of prior exposures of the current ad ($A_{i,t}$) within the session, and (2) $\mathit{Space}_{i,t}$ – temporal space between the current ad and the last time it was shown in the session. For example, if session i shows the sequence $\langle A, A, B, C, A \rangle$, we have $\mathit{Freq}_{i,5} = 2$ and $\mathit{Space}_{i,5} = 5 - 2 = 3$.[11] While the qualitative results on the coefficient of variety do not change, the significance level of coefficients for both $\mathit{Freq}_{i,t}$ and $\mathit{Space}_{i,t}$ highlights the importance of controlling for other session-level variables.[12]

Overall, our preliminary results in Table 2 show a strong statistical link between the variety of previous ads and the click outcome on the current ad. The patterns are quite robust: we find the same patterns when we use an entropy-based measure of variety such as Shannon Entropy, as well as when we use a logistic regression to estimate our binary outcome (please see Web Appendix B for the results). However, we must be careful to read these results as anything more than preliminary evidence for the effects of variety.

Challenges in Establishing Causality

We now discuss why our preliminary analysis falls short of establishing a causal link between the variety of previous ads and the click outcome on the current ad. These reasons relate to three different aspects of our main variable $V_{i,t}$.

First, as illustrated in Figure 5b, assignment to variety is not fully exogenous. While the model in the second column of Table 2 tries to address this issue by controlling for targeting variables through fixed effects for each targeting sub-category, there may still be a more complicated selection (e.g., through the interaction of different targeting variables). Thus, it is essential for our empirical framework to guarantee that, conditional on controls, the assignment to variety is fully exogenous.

The second issue stems from the fact that the receipt of variety is different from assignment to it. This is because users may leave the session at any point they want. The discrepancy between the receipt and assignment can create identification challenges because we only observe the receipt, whereas randomization (if any) happens at the assignment level. Notice that the control for the

[11] If the ad has not been shown in the session before, we have $\mathit{Space}_{i,t} = t$

[12] These variables are correlated with variety as follows: $\rho(V_{i,t}, \mathit{Freq}_{i,t}) = -0.1702$ and $\rho(V_{i,t}, \mathit{Space}_{i,t}) = 0.4168$.

session length in the third column of Table 2 only helps in cases where the users decide how long they want to stay in a session before it starts. However, there may be more complex scenarios wherein users' decision to leave is a function of their variety assignment. Thus, our empirical framework needs to account for the discrepancy between the receipt and assignment to variety.

Finally, variety is not very well-defined as a treatment. It is hard to isolate an exogenous increment in variety such that only variety changes one unit, with all else remaining constant. As such, even with complete randomization of variety and no dynamic selection, the models in Table 2 estimate a composite effect of variety and other session-level variables that co-vary with variety. Controlling for other session-level variables (Column 4 in Table 2) helps with this issue, but we cannot verify whether these controls are sufficient. Thus, a primary goal in our empirical framework is to isolate the effects of variety to the extent possible.

In sum, these three issues preclude us from making causal statements based on our preliminary analysis. In the next section, we discuss our empirical strategy to address these issues.

Empirical Framework

In this section, we present our empirical framework. We start by defining the problem. Next, we formally discuss the selection issues and present our identification strategy to address these issues. Finally, we present our full model specification and estimation method.

Problem Definition

Before we formally define our problem, we re-state the goal of our study: we want to examine the extent to which a random increment in the variety of previous ads changes the click outcome, holding everything else constant. As highlighted in our preliminary analysis, a fundamental challenge in achieving our goal is the difficulty in randomizing this increment in isolation: an increment in variety may change other session-level covariates, thereby violating the ceteris paribus interpretation we are aiming at. For example, an increment in the variety of previous ads can change the spacing between ads as well. Thus, this challenge boils down to whether we can achieve some level of separability between an increment in variety and the rest of the information in the prior sequence of ads shown such that we can manipulate variety, holding all else constant.

An interesting feature of our measure of variety that helps with this separability condition is the fact that an increment in this measure has a clear interpretation at the exposure level: every exposure that shows an ad that has not been shown before adds one unit to the breadth of variety, thereby capturing the event of "increase in ad variety". We can formalize this intuition by defining the binary variable $W_{i,t}$ for any $t \geq 3$ as follows:

$$W_{i,t} = \mathbb{1}(A_{i,t-1} \notin \{A_{i,1}, \ldots, A_{i,t-2}\}) \tag{4}$$

Hence, $W_{i,t}$ takes the value one if the ad shown in exposure $t-1$ is distinct from the set of ads shown in the prior $t-2$ exposures, thereby increasing ad variety by one unit.[13] This allows us to derive a binary decomposition of variety at the exposure level for any $t \geq 3$, as follows:

$$\begin{aligned}
V_{i,t} =& |\{A_{i,1}, \ldots, A_{i,t-1}\}| \\
=& |\{A_{i,1}, \ldots, A_{i,t-2}\}| + \mathbb{1}(A_{i,t-1} \notin \{A_{i,1}, \ldots, A_{i,t-2}\}) \\
=& V_{i,t-1} + W_{i,t} \\
=& 1 + \sum_{s=3}^{t} W_{i,s}.
\end{aligned} \tag{5}$$

The intuition behind this decomposition is simple: for each distinct ad in the set of prior ads, $W_{i,t}$ takes value one only once (the first time each distinct ad is shown in the sequence).

The recursive relationship $V_{i,t} = V_{i,t-1} + W_{i,t}$ in the equation above illustrates how we approach separability in our problem: focusing on the increment in the last exposure helps us isolate its effects from the sequence of ads shown prior to that. As such, at any exposure $t \geq 3$, we define the binary variable $W_{i,t}$ as the treatment variable of interest. An intuitive definition of $W_{i,t}$ is "an increase in ad variety" in the previous exposure. Our goal would then be to measure the effect of this treatment on

[13] For example, in a session i with ads $\langle A, B, A, C \rangle$, we have $W_{i,3} = 1$ and $W_{i,4} = 0$.

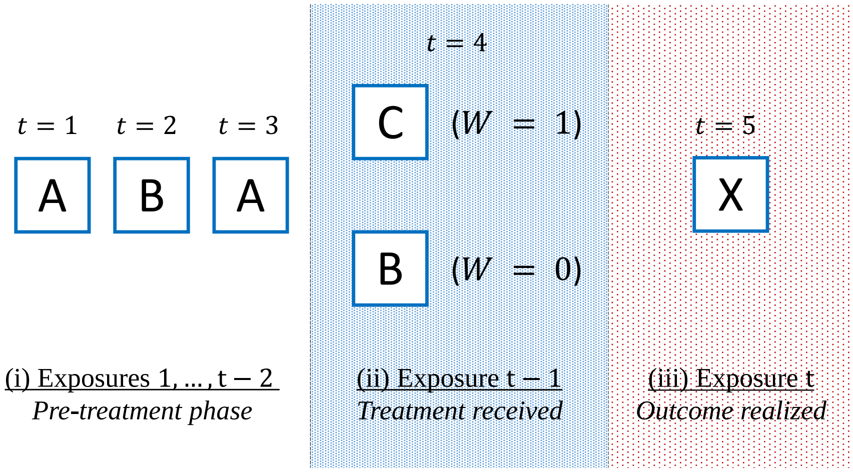


Figure 6: A visual depiction of our research design. There are three phases – (i) pre-treatment period, which includes all the information about the user as well as all the ads seen up until exposure $t-1$, (ii) treatment period in exposure $t-1$, where the user is assigned to the variety condition or control, and (iii) outcome collection period t, where we collect the click outcome.

the click outcome in the current exposure. Figure 6 visualizes our research design, where the variety treatment is assigned and received in exposure $t-1$ and the outcome will be collected in exposure t.[14] This research design helps us overcome/avoid the two shortcomings in our preliminary analysis: (1) the receipt of variety is the same as the assignment to it, and (2) the treatment increment has a clearer definition, as in, it does not co-vary with the session-level information up until the treatment stage (pre-treatment period in Figure 6).

With the goal of measuring the causal effects of "an increase in ad variety at point $t-1$" on the "click outcome at point t" in mind, we write the main equation we want to estimate as:

$$\begin{aligned} Y_{i,t} =& \beta W_{i,t} + g(X_{i,t}, H_{i,t}) + \epsilon_{i,t}, \\ =& \beta W_{i,t} + \underbrace{g_{pre}(X_{i,t-2}, H_{i,t-2})}_{\text{pre-treatment controls}} + \underbrace{g_{post}(A_{i,t-1}, A_{i,t}; H_{i,t-2})}_{\text{post-treatment confounding}} + \epsilon_{i,t}, \end{aligned} \tag{6}$$

where β captures the effects of variety and the second equality represents our use of separability

[14] A natural question is why we do not compare the outcomes at point $t-1$. The main reason is that it would not be possible to separate the variety effects from ad effects without assuming a certain specification. It is worth noting that under these specifications, the results are directionally the same as our main results. In the current research design, however, we have greater power to control for the ad effects in the period after treatment assignment.

in this problem. We explicitly assume that we can additively separate our controls into two categories: (1) pre-treatment controls that account for selection in the assignment to our treatment, i.e., $g_{pre}(X_{i,t-2}, H_{i,t-2})$, (2) post-treatment controls that help separate the treatment effects from other post-treatment variables that co-vary with our treatment, i.e., $g_{post}(A_{i,t-1}, A_{i,t}; H_{i,t-2})$.

Before we specify the functions g_{pre}, g_{post}, we first describe the challenges that we face and our identification strategy to address these challenges. Consistent with our identification strategy, we will then present our full model specification.

Identification Strategy

We face three key challenges in estimating the treatment effects specified in Equation (6):

- *Pre-treatment confounding:* Assignment to treatment is confounded by pre-treatment variables. That is, a user's propensity to receive the treatment is a function of these pre-treatment variables. Function g_{pre} in Equation (6) should address this type of confounding in our problem.
- *Dynamic selection (post-treatment censoring):* While the assignment to variety is equivalent to its receipt, the user may leave right after receiving the treatment (i.e., just after the $t-1$th exposure) thereby censoring some of the post-treatment variables and the outcome. We need to address this dynamic selection problem in our identification strategy.
- *Post-treatment confounding:* The post-treatment phase is defined from the treatment assignment to the outcome collection phase. During this time, other important variables also co-vary with our treatment (e.g., the identity of the ads shown, their recency, and frequency). This imposes a challenge if we want to isolate the treatment effect of an increase in ad variety. Therefore, we need to control for any post-treatment confounding to isolate the effects of variety. Function g_{post} in Equation (6) should control for this type of confounding.

Solution to Pre-treatment Confounding

The variation in our treatment variable $W_{i,t}$ can be confounded with the pre-treatment variables. As shown in Figure 3c, advertisers favor more popular sessions whose characteristics are associated with higher CTR. As a result, assignment to treatment is more likely in more popular sessions simply because there are more ads in the inventory to increase the variety of prior sequence. For instance, in our data, over 83% of the impressions in the largest province are assigned to the treatment condition, whereas this percentage drops to 61% for impressions in a small province. Further, as shown in Figure 6, the assignment to treatment at exposure $t-1$ depends on the prior ad assignments in the session at exposures 1 to $t-2$. For example, if the variety of prior ads in a session is higher, the likelihood of being assigned to the treatment is lower by construction. For example, 76% of impressions whose prior variety was one received the treatment, whereas only 40% of impressions with prior variety of four received the treatment. Thus, we can state this challenge as follows:

Challenge 1. *There is non-randomness in the treatment assignment. That is, for an arbitrary exposure t in two random sessions i and j, the propensities of receiving the treatment are not necessarily the same, i.e.,* $\Pr(W_{i,t}=1) \neq \Pr(W_{j,t}=1)$.

To solve this problem, we focus on the source of non-randomness in the treatment variable. Given the definition of our treatment variable in Equation (4), we know that the distribution of the ad allocation process fully determines the distribution of treatment assignment. Hence, we focus on the ad allocation process as the source of non-randomness in the treatment. In light of Equation (1), we know that advertisers' bids, quality scores, and participation decisions fully determine the ad allocation process. Therefore, these are the only three possible sources of non-randomness. We use this observation to formally show the following proposition:

Proposition 1. *The distribution of propensity scores for ad assignment, $\pi_{i,t}(a)$s, for any exposure/impression is only a function of impression-specific observables, $X_{i,t}$, in the data.*

Proof. See Web Appendix C.1. □

This is a crucial result since it ensures, conditional on exposure, that there are no user or impression-specific unobservables that affect ad assignment that are observable to advertisers (and the ad-network) but not to the researcher. We now link Proposition 1 to the propensity scores for our treatment variable $W_{i,t}$ in the following remark:

Remark 1. *Let* $e(W_{i,t})$ *denote the propensity score to be assigned to the treatment condition, i.e.,* $e(W_{i,t}) = \Pr(W_{i,t} = 1)$*. Then* $e(W_{i,t})$ *is only a function of impression-specific observables, since* $e(W_{i,t})$ *is a linear function of* $\pi_{i,t}(a)$*s at any point, i.e.,* $e(W_{i,t}) = \sum_{a \notin H_{i,t-2}} \pi_{i,t-1}(a)$*.*

The fact that $e(W_{i,t})$ is only a function of impression-specific observables is important because it shows the unconfoundedness of our treatment variable $W_{i,t}$. That is, for any set of potential outcomes $\mathcal{Y}_{i,t}$, we have $\Pr(W_{i,t} \mid X_{i,t-1}, H_{i,t-2}) = \Pr(W_{i,t} \mid \mathcal{Y}_{i,t}, X_{i,t-1}, H_{i,t-2})$.

Our approach to directly address Challenge 1 is to use one of the key results of Rosenbaum and Rubin (1983) – it is sufficient to control for the propensity scores of the treatment variable ($e(W_{i,t})$) under unconfoundedness,. As such, the challenge boils down to estimating the propensity scores of the treatment variable using the pre-treatment information. Estimation of $e(W_{i,t})$ is a prediction task where we need to regress observed $W_{i,t}$ in the data on $X_{i,t-1}$ and $H_{i,t-2}$. Since this is a prediction task we can use a machine learning method that can capture more complex relationships and achieve better predictive accuracy. Our specific solution to address the pre-treatment confounding challenge is summarized in three steps as follows:

- *Step 1:* We estimate $e(W_{i,t})$s using an XGBoost model (Chen and Guestrin, 2016). That is, $\hat{e}(W_{i,t}) = \hat{XGB}(X_{i,t-1}, H_{i,t-2})$. Please see Web Appendix §D.1 for the details of our propensity estimation approach.
- *Step 2:* We assess covariate balance to confirm that inverse propensity weight-adjusted distribution of each variable is reasonably similar across treatment and control groups. Please see Web Appendix §D.2 for the details of our balance assessment.
- *Step 3:* We feed the inverse propensity weights to the regression model to account for the pre-treatment confounding issue.

Solution to Dynamic Selection

Next, we discuss the issue of dynamic selection. As discussed earlier, users' decision to leave a session can cause a missing data problem. In our setting, dynamic selection only happens at the outcome level: while all the users available at point $t-1$ receive the variety treatment, their outcome

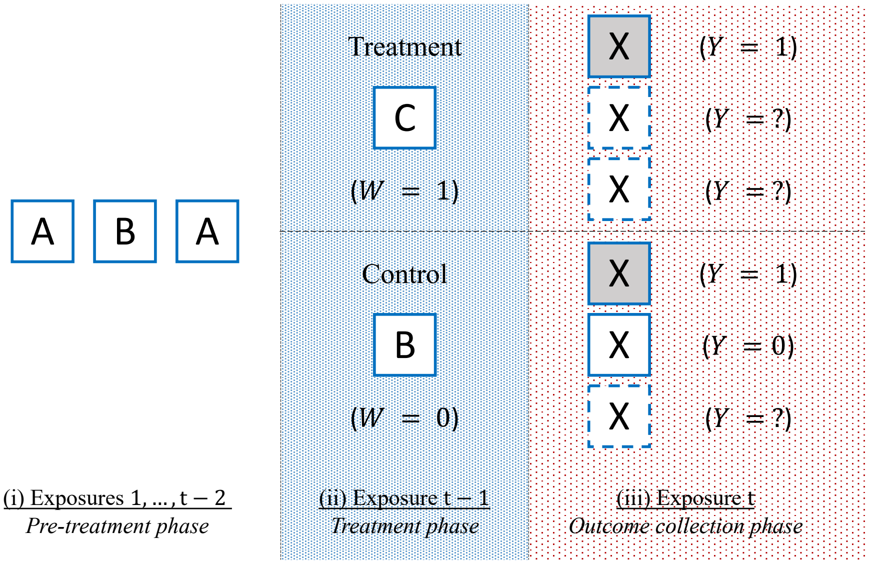


Figure 7: Example of post-treatment censoring. The dotted impressions are exposures that did not occur because the user left the session prior to their occurrence and the grey shaded impressions denote clicks.

at exposure t may be missing if they decide to leave the session right after the treatment. This type of dynamic selection does not create estimation bias if users' decision to leave is completely random. However, dynamic selection can interfere with inference if a user's decision to leave the session is: (1) a function of her variety assignment in exposure $t-1$, or (2) her own characteristics (especially unobserved) that also affect her click probability.[15]

Figure 7 uses a simple example to illustrate the general form of dynamic selection in our problem and how it can interfere with inference. In this example, there are six users who have seen ads $\langle A, B, A\rangle$ in their first three exposures. These users are then randomly assigned to treatment and control conditions in the treatment state ($t = 4$). In all these cases, the user is supposed to see ad X in the fifth exposure. However, only three out of six were exposed to the fifth exposure: one in the treatment condition and two in the control condition. To see how dynamic selection can interfere with inference, suppose that only one impression in each condition has been clicked. If we observe all the exposures (i.e., dynamic selection is not an issue), then we would correctly infer that there is

[15] Dynamic selection has also been recognized in other settings, where users' survival is a function of user-level unobservables. For example, Yoganarasimhan (2013) accounts for persistent user-level unobservables within dynamic structural models by explicitly incorporating unobserved heterogeneity in users' utility functions and state transitions.

no difference in CTRs across the treatment and control conditions. However, if we solely rely on the observed data, we would infer a CTR of 1 for the treatment condition and a CTR of $\frac{1}{2}$ for the control condition. This would lead to the incorrect inference that treatment results in a higher CTR.

To address this issue, we need to impute the missing impressions. In particular, we need to impute two characteristics of these missing impressions – (1) the ad the user would have been assigned to, and (2) the corresponding click outcome. Let $A^*_{i,t}$ and $Y^*_{i,t}$ denote the ad and click outcome for both observed and missing impressions. For observed impressions, we have $A^*_{i,t} = A_{i,t}$ and $Y^*_{i,t} = Y_{i,t}$. We can now formally state this challenge as follows:

Challenge 2. *For a user who has received the treatment at exposure $t - 1$ but did not stay for exposure t, we need to impute the set of $\{A^*_{i,t}, Y^*_{i,t}\}$, where $A^*_{i,t}$ is the ad this user would have been assigned to in exposure t, and $Y^*_{i,t}$ is the corresponding outcome.*

We first discuss our imputation strategy for missing ads. Let τ denote the exact time-stamp of an exposure in session i.[16] Next, let the the ad shown in session i at timestamp τ be drawn from the distribution $\mathcal{A}_i(\tau)$. Then, it is easy to show the following remark:

Proposition 2. *For any two exposures in sessions i and j with the same targeting characteristics, the distribution of ad allocation is the same at any arbitrary timestamp τ, i.e., $\mathcal{A}_i(\tau) \equiv \mathcal{A}_j(\tau)$.*

Proof. See Web Appendix C.2. □

Proposition 2 is a direct result of the ad allocation process in Equation (1). In light of this proposition, we can use the actual ad assignment in exposures from other sessions that are not part of our sample (but share the same targeting characteristics) to impute the intended ad assignment for exposures in the sessions in our sample.[17] We present the details of our imputation approach in Web Appendix §E. It is worth noting that unlike most imputation approaches that use models to approximate the original distribution and simulate missing data from this approximate distribution,

[16] Note that timestamp τ is distinct from exposure number t; τ is the exact time at which an impression occurs, e.g., if the first impression in a session occurred at 9:21:34 pm of a specific day, then τ is 9:21:34 pm whereas $t = 1$.

[17] We do not use the sessions in our sample for imputation because doing so can cause finite-sample issues in some parts of the data.

our approach is model-free and guarantees that the imputed exposures are drawn from the original distribution (e.g., $A^*_{i,T_i+1} \sim \mathcal{A}_i(\tau)$ in the example above).[18]

Finally, we impute the missing outcomes as zero simply because the user is not available to click on the ad. An alternative approach would be to impute the outcome as the click decision the user would have made had she stayed in the session (Little and Yau, 1996). While this is the conventional approach in medical studies, we believe that our approach is the right one for our context because the leave decision by the user prevents the event where the outcome of interest happens (the user clicks on the next ad).[19] Nevertheless, we run a series of robustness checks to show our results are not driven by this modeling choice; see Web Appendix §F.4 for details.

Solution to Post-treatment Confounding

While unconfoundedness rules out pre-treatment confounding, the nature of our treatment gives rise to the issue of post-treatment confounding. That is, from the point users are assigned to the treatment (at exposure $t-1$) to the point we collect the outcome, it is not just the treatment assignment that is different across treatment and control groups: other information regarding exposures $t-1$ and t may also differ, such as the specific ads shown as well as their prior frequency and spacing in these exposures. This brings us to our third challenge, which we summarize as follows:

Challenge 3. *There exists a function $g_{\text{post}}(A_{i,t-1}, A_{i,t}; H_{i,t-2})$ that is defined on post-treatment inputs and is correlated with $W_{i,t}$ as well as $Y_{i,t}$. Therefore, failure to control for this function leads to omitted variable bias.*

To address this challenge, we need to specify function g_{post} such that it captures any post-treatment variable in $\langle A_{i,t-1}, A_{i,t}; H_{i,t}\rangle$ that is correlated with both $W_{i,t}$ and $Y_{i,t}$. To simplify the problem, we need to define variables that capture the relationship between $A_{i,t}$ and $A_{i,t-1}$ with the past sequence $H_{i,t-2}$. We focus on two sequence-dependent variables that have been shown to drive

[18] While we use this specific approach to overcome dynamic selection, the general solution is to approximate the allocation distribution from the data, which is feasible under the unconfoundedness of ad allocation.

[19] This is different from medical studies where the outcome is often an objective measure of one's health rather than a choice.

ad effects – frequency and spacing of the ad shown in an exposure in the session. For any exposure t, we define the within-session frequency of the ad shown at this exposure, $\mathit{Freq}_{i,t}$, as:

$$\mathit{Freq}_{i,t} = \sum_{s=1}^{t-1} \mathbb{1}(A_{i,s} = A_{i,t}) \tag{7}$$

This variable simply measures the number of times the ad shown at exposure t has been shown earlier in the session. Next, we define $\mathit{Space}_{i,t}$ as the spacing between the ad shown at exposure t and the last time this ad has been shown (if any). We can write:

$$\mathit{Space}_{i,t} = t - \max\{s \cup \{0\} \mid A_{i,s} = A_{i,t}\}, \tag{8}$$

where the spacing is defined in terms of exposure numbers and we have $\mathit{Space}_{i,t} = t$ if the ad shown at t has not been shown before. We use these two variables to eliminate the dependence on the past. That is, we assume that all the information in $\langle A_{i,t-1}, A_{i,t}; H_{i,t}\rangle$ is summarized in $\langle A_{i,t-1}, \mathit{Freq}_{i,t-1}, \mathit{Space}_{i,t-1}, A_{i,t}, \mathit{Freq}_{i,t}, \mathit{Space}_{i,t}\rangle$. In light of this assumption, we need to include all these six variables that are not perfectly co-linear with our treatment $W_{i,t}$. We notice that both $\mathit{Freq}_{i,t-1}$ and $\mathit{Space}_{i,t-1}$ are perfectly co-linear with our treatment:

$$W_{i,t} = \mathbb{1}(A_{i,t-1} \notin |\{A_{i,1}, \ldots, A_{i,t-2}\}|) = \mathbb{1}(\mathit{Freq}_{i,t-1} = 0) = \mathbb{1}(\mathit{Space}_{i,t-1} = t-1) \tag{9}$$

Therefore, we exclude these two variables from our model. However, we will come back to these two variables when exploring the potential behavioral mechanisms.

In sum, we include the following four variables to control for post-treatment confounding – $A_{i,t-1}$, $A_{i,t}$, $\mathit{Freq}_{i,t}$, and $\mathit{Space}_{i,t}$. Together, our post-treatment controls allow us to isolate the effects of treatment to the maximum extent possible. Later in Web Appendix §F.2, we conduct a series of robustness checks to confirm that our findings are robust to alternative estimation approaches.

Model Specification

We now discuss our model specification based on the identification strategy for three main challenges. To address the pre-treatment confounding issue, we use IPW-adjusted linear regression where we weight impressions by their inverse propensity score of receiving treatment, since it is an efficient estimator when using propensity scores to estimate treatment effects (Hirano et al., 2003). Next, to account for dynamic selection, we run our regression using the fully imputed variables for the ad and click outcome, i.e., $Y^*_{i,t}$ and $A^*_{i,t}$. Finally, we address post-treatment confounding by controlling for $A_{i,t-1}$, $A_{i,t}$, $\mathit{Freq}_{i,t}$, and $\mathit{Space}_{i,t}$. Thus, the main version of our model is the following IPW-adjusted regression specification:

$$Y^*_{i,t} = \beta W_{i,t} + \sum_q \gamma_q \mathbb{1}(\mathit{Freq}_{i,t} = q) + \sum_s \delta_s \mathbb{1}(\mathit{Space}_{i,t} = s) + \alpha_0(A^*_{i,t}) + \alpha_1(A_{i,t-1}) + \zeta_t + \epsilon_{i,t}, \quad (10)$$

where β captures the treatment effect, γ_q and δ_s are the coefficients for levels q and s of $\mathit{Freq}_{i,t}$ and $\mathit{Space}_{i,t}$ respectively, $\alpha_0(A^*_{i,t})$ and $\alpha_1(A_{i,t-1})$ control for the fixed effects of ads shown in exposures t and $t-1$, and ζ_t controls for exposure number fixed effects. We use this model as the main specification, but we also consider other specifications with more controls in the next section.

Results

We now present our results. We first establish the average effects of variety. Next, we present a series of robustness checks on main findings of the paper. Finally, we delve deeper into the underlying mechanism and empirically test the predictions of the mechanism.

Main Effects of Variety

Results from the Main Specification

We start by estimating the average treatment effect for our main specification in Equation (10), and present the results in the first column of Table 3. We use the sample of impressions from exposures $t = 4$ to $t = 10$.[20] Since we use IPW-adjusted regression, we use robust standard errors

[20] This is because we want to start from an exposure that has a relatively high propensity of assignment for both treatment and control groups. For example, starting from $t = 2$ gives us very low propensity scores for the control condition.

for inference. The positive and significant treatment coefficient indicates a positive causal link between an increase in the variety of prior ads and the click outcome. That is, showing an ad that increases the variety of the sequence of ads results in a higher click-through rate on the next ad, holding all else constant. Our main finding highlights a source of externality in this market – an intervention in a given period affects the outcomes in future periods. This is in contrast with the common assumption in the online advertising marketplace, where each impression is sold as an independent unit.[21]

To interpret the magnitude of our coefficients, we compare them with the baseline CTR in the system. The baseline CTR for our sample from $t = 4$ to $t = 10$ is 0.0134, which means that around 1.34% of the impressions in our sample get clicked. The treatment coefficient in the first column of Table 3 is 0.00186. As such, the magnitude of our treatment coefficient accounts for 13.88% of the baseline CTR in our sample. This implies that an increase in ad variety can shift CTR by roughly 13.88%, holding all other variables fixed.

Next, we try other specifications by dropping some of the necessary controls. First, in the second column of Table 3, we run an unweighted least squares regression to compare the estimates with and without IPW adjustment. As expected, both the magnitude and statistical significance increase since the unweighted model does not account for the differences in propensity scores, thereby capturing the endogenous variation in treatment assignment. In the third column, we run the model without accounting for the dynamic selection issue. That is, we only focus on the sample of impressions that survived, and drop the impressions where the user left the session right after the treatment assignments. Notice that the outcome for all the excluded impressions is zero. As a result, the estimated treatment coefficient is therefore not directly comparable to the other coefficients in Table 3 because the samples are systematically different. To adjust for that, we need

For the ending point, we want to end at a t that still has enough impressions. It is worth emphasizing that our results do not change if we include all time periods.

[21] The common practice of running second- or first-price auctions to sell digital ads is based on the assumption that impressions are independent units.

	Dependent variable: Click ($Y^*_{i,t}$)			
	(1)	(2)	(3)	(4)
Treatment ($W_{i,t}$)	0.00186*** (9.01)	0.00203*** (11.14)	0.00247*** (9.91)	0.00235*** (11.81)
IPW Adjustment	✓		✓	✓
Imputed Sample	✓	✓		✓
Exposure (t) FE	✓	✓	✓	✓
Freq$_{i,t}$ Indicators	✓	✓	✓	
Space$_{i,t}$ Indicators	✓	✓	✓	
$A^*_{i,t}$ FE	✓	✓	✓	✓
$A_{i,t-1}$ FE	✓	✓	✓	✓
No. of Obs.	2,405,695	2,405,695	1,993,542	2,405,695
R^2	0.006	0.006	0.007	0.005
Adjusted R^2	0.005	0.006	0.007	0.005
Note:	*p<0.05; **p<0.01; ***p<0.001			

Table 3: Average effects of the variety treatment on the CTR. Numbers reported in parentheses are t-statistics computed based on robust standard errors.

to multiply the coefficient in the third column by the ratio of two samples 1,993,542/2,405,695, which is 0.8287. The adjusted coefficient is 0.00205, which implies that we would overestimate the effects of treatment if we do not account for dynamic selection caused by the effect of treatment on users' decision to leave. This is because an increase in ad variety likely comes with a greater user propensity to leave the session, as an increase in ad variety can be perceived as increased ad load, which has been shown to have negative effects on usage in the prior advertising literature (Wilbur, 2008). Finally, in the fourth column of Table 3, we drop the controls, *Freq*$_{i,t}$ and *Space*$_{i,t}$. Given that both these covariates are correlated with our treatment and likely associated with the click outcome as well, we expect a change in the treatment effect estimates. We find that we would overestimate the treatment effects if we do not control for these two covariates.

In sum, the results in Table 3 establish the main positive effects of an increase in ad variety on the click outcome on the next ad, and shows the importance of controlling for all three types of confounding discussed in the paper.

Robustness Checks on the Main Effects

We perform a series of robustness checks on our main results. We discuss these models briefly here, and refer readers to Web Appendix §F for details.

First, in Table 3, we user a linear probability model for a binary outcome. In the Web Appendix §F.1, we present the results from a logistic regression for the same model specifications. Second, we consider models with overly conservative controls for both pre- and post-treatment variables. These models separately control for: (1) interaction of all targeting variables[22], (2) user and hour-day fixed effects, (3) session fixed effects, and (4) different interactions of all post-treatment variables. Overall, as shown in the Web Appendix §F.2, our results consistently show the main effect: that an increase in ad variety at any exposure results in higher CTR to the next ad.

Most notably, we employ an exact matching approach to fully isolate our treatment effects from any pre- or post-treatment covariates. In this practice, we match impressions based on the exact sequence of ads shown in the session except the ad shown at the treatment phase (exposure $t-1$). That is, two impressions belong to the same matching group if $\langle A_{i,1}, \ldots, A_{i,t-2}, A_{i,t} \rangle$ is exactly the same for them and they only differ in $A_{i,t-1}$. We further control for the fixed effects of $A_{i,t-1}$ and the propensity scores of the treatment. Even though our statistical power is substantially compromised in this case, our main findings still hold and the results of this exact matching approach show a significant and positive treatment coefficient. Please see Web Appendix §F.3 for more details on our exact matching practice.

Finally, we run a series of additional checks to establish the robustness of our results to (1) alternative approaches to imputation; see Web Appendix §F.4 for details, (2) different levels of clustering in standard errors; see Web Appendix §F.5 for details, and (3) a placebo treatment definition to ensure the data structure does not drive our results; see Web Appendix §F.6 for details. In sum, all these robustness checks confirm the validity of our main results.

[22] This approach is similar to Nair et al. (2017) who use firms' targeting decision to control for the selection caused by targeting. While our main approach is based on using propensity scores, here we add the interaction of all targeting variables to make sure that our main results are not driven by our propensity score estimates.

Mechanism for the Effects of Variety

We now examine the mechanism behind the main results and present empirical evidence in support of our ideas.

Theoretical Underpinnings of the Mechanism

In the previous section, we found that showing a new (or previously unseen) ad at $t-1$ increases the user's probability of clicking on the ad shown at t. To pin down the mechanism, we focus on the main feature that differs across our treatment and control groups – the novelty of the ad shown at the treatment phase. A unifying result that emerges from both early and recent work in the behavioral literature is that novelty of stimuli shown in a given space increases subjects' attention to that space (Helson, 1948; Kahneman, 1973; Han and Marois, 2014). In our context, this means that the increase in ad variety by showing a novel ad leads to higher attention to the advertising slot, thereby increasing the consideration and click probability on the next ad. We propose this attention-based account as the underlying mechanism for the variety effects found in the paper.

There are three key advantages in using the aforementioned attention-based account. First, it is parsimonious as it only uses a well-established finding that users pay more attention to novel stimuli (ads in this case) that have been used less recently and less frequently in the past that traces its roots to the early work on adaptation-level theory (Helson, 1948), which has been a theoretical foundation to study the effects of variety (Redden, 2008). Second, the foundation of our theory, that users pay more attention to novel stimuli, is consistent with earlier papers in advertising and eye-tracking (Pieters et al., 1999), thereby giving us contextual validity. Finally, our behavioral account delivers concrete predictions that are empirically testable: we can focus on different slices of the data where we expect to have a higher (or lower) gap in novelty between the treatment and control, and test if the estimates consistently change with this gap in novelty. In the rest of this section, we adopt this strategy, i.e., we make theory-driven predictions based on our proposed mechanism and then examine if they are empirically true.

Treatment Effects Across Different Control Groups

Recall Equation (9) in the paper:

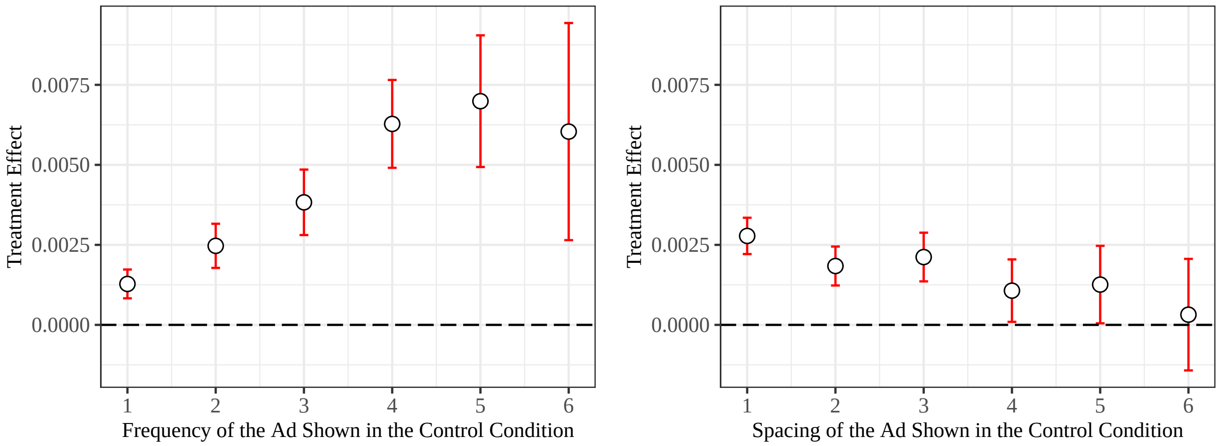
$$W_{i,t} = \mathbb{1}(A_{i,t-1} \notin |\{A_{i,1}, \ldots, A_{i,t-2}\}|) = \mathbb{1}(\textit{Freq}_{i,t-1} = 0) = \mathbb{1}(\textit{Space}_{i,t-1} = t-1)$$

In light of this equation, we know that an impression belongs to the treatment condition if the ad shown at period $t-1$ has not been shown in the past, i.e., $\textit{Freq}_{i,t-1} = 0$ and $\textit{Space}_{i,t-1} = t-1$. On the other hand, an impression belongs to the control condition if $\textit{Freq}_{i,t-1} \neq 0$ and/or $\textit{Space}_{i,t-1} \neq t-1$. In other words, our control condition constitutes a range of values for $\textit{Freq}_{i,t-1}$ and/or $\textit{Space}_{i,t-1}$. However, our underlying mechanism suggests that the exact levels of past frequency and spacing of the control condition matter. Specifically, we expect the control group to be less novel if the ad at the treatment phase has been shown more frequently (higher $\textit{Freq}_{i,t-1}$) and/or more recently (lower $\textit{Space}_{i,t-1}$). Building on these ideas, we can formulate the following concrete predictions:

Prediction 1. *The treatment effect is higher when we compare the treatment group ($W_{i,t} = 1$) to a control group where the ad shown at the treatment phase has higher frequency. That is, as we increase $k > 0$ in* $\text{Freq}_{i,t-1} = k$ *to define the control group, the treatment effect increases.*

Prediction 2. *The treatment effect is higher when we compare the treatment group ($W_{i,t} = 1$) to the control group where the ad shown at the treatment phase has lower spacing (higher recency). That is, as we increase $l < t-1$ in* $\text{Space}_{i,t-1} = l$ *to define the control group, the treatment effect decreases.*

To empirically test these predictions, we first partition the control group in our data into subgroups based on the frequency of the ad shown at $t-1$ (e.g., $W_{i,t} = 0$ and $A_{i,t-1}$ has been shown k times before). Then, we separately estimate the treatment effect against each of these control groups and present the results in Figure 8a. We see an increasing trend in treatment effects as the frequency of the ad at $t-1$ increases. This suggests that when the last ad shown before t (i.e., $A_{i,t-1}$) was repeated many times earlier in the session (i.e., is less novel), users pay less attention to the current



(a) Control with Different Frequency

(b) Control with Different Spacing

Figure 8: Treatment effects when compared with different control groups defined by the frequency ($Freq_{t-1}$) and spacing ($Space_{t-1}$) of the ad shown in the control condition. Confidence intervals are built using the robust standard errors from the IPW-adjusted regression model.

ad. We perform a similar exercise by partitioning the control group into different subgroups based on recency or spacing (e.g., $W_{i,t} = 0$ and $A_{i,t-1}$ has been shown l exposures before) and present the results in Figure 8b. Here, we see a decreasing pattern – the treatment effects are decreasing as the spacing of the ad shown in the control condition increases. The highest treatment effect is when the treatment condition is compared with the control condition that repeats the ad before ($A_{i,t-1} = A_{i,t-2}$), which is equivalent to spacing level of one). Interestingly, when we increase the spacing level to six in the control condition (i.e., $A_{i,t-1}$ was shown 6 impressions prior to $t - 1$), there is no significant treatment effect anymore. This suggests that repeating an ad that was shown much earlier in the session is almost as good as showing a new ad (i.e., the treatment condition where the variety increases by one unit). Together, the above findings provide support for our proposed mechanism. See Web Appendix §G.1 for the details of the regression used in Figure 8 and additional robustness checks on these findings.

Heterogeneity Across Usage Frequency and Recency

To further explore the idea that users' response to the variety treatment is driven by stimulus-novelty and attention, we focus on two user-level features that capture the user's pre-session exposure to ads: (1) *usage frequency* – the number of ad impressions the user has seen in prior sessions, and (2)

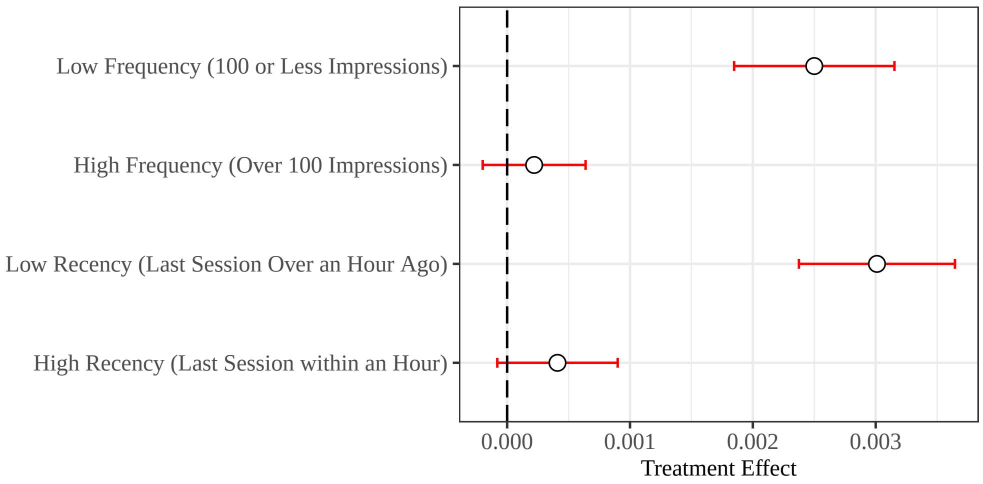


Figure 9: Heterogeneity in the variety effects across usage frequency and recency. Confidence intervals are built using robust standard errors in the IPW-adjusted regression.

usage recency, which denotes the time lapse between the start of the current session and the end of her previous session. This variable captures how long ago the user was exposed to an ad (before the currrent session): when this gap is short, the usage recency is high.

The theoretical mechanism proposed earlier would suggest that it is harder to shift users' attention if they have seen more ads in the past. Hence, we expect to see higher treatment effects for users with lower frequency of prior ads. Similarly, users who had more recent interactions with ads are less likely to be responsive to our treatment because their memory of some ads can be fresh/recent. As such, we expect the within-session interventions to be less effective in shifting users' attention when they are in the high recency condition. Together, we can form the following two predictions based on our mechanism:

Prediction 3. *The treatment effect is higher for users with low usage frequency as compared to users with high usage frequency.*

Prediction 4. *The treatment effect is higher for users with low usage recency as compared to users with high usage recency.*

We now empirically test these predictions in our data. We perform a rough median split and define the high (low) usage frequency sample as the set of impressions where the user has seen over (less than) 100 impressions in prior sessions. We then estimate the treatment effects separately for these two partitions of the data. Next, we perform a similar exercise for usage recency, where the one hour is a rough median split. We then use one hour as the threshold to partition impressions into low and high recency buckets, and estimate separate treatment effects for each bucket. The estimated treatment effects from these analyses are shown in Figure 9. Notice that the treatment effects are significant and positive only in low usage frequency or recency conditions. When the usage frequency or recency is high, the treatment is statistically insignificant.[23] These findings are consistent with Predictions 3 and 4, which are based on the attention-based mechanism proposed earlier.

Heterogeneity Across Past Variety

We now examine the heterogeneity in our treatment effects across past variety. In light of our mechanism, we expect the treatment to be less effective if the prior variety is already high. This is because the increment in attention from treatment assignment would be lower when attention level is already high due to a higher past variety. We can state this prediction as follows:

Prediction 5. *The treatment effect is lower when the variety of $\langle A_{i,1}, \ldots, A_{i,t-2} \rangle$ is higher.*

We test this prediction by estimating models that also include an interaction of the treatment with past variety. We consider three different measures of past variety in our analyses: (1) $V_{i,t-1}$, which is simply the number of distinct ads shown in the first $t-2$ exposures, (2) $\log(V_{i,t-1})$, which is the logarithm of the previous measure, and (3) $V_{i,t-1}/(t-1)$, which normalizes our first measure across exposure numbers. The results from this exercise are shown in Table 4. Across all specifications and definitions of prior variety, our interaction term is negative and statistically significant. This implies that the treatment effect reduces as past variety increases, i.e., a novel ad (or the variety treatment) is less likely to increase user attention when their attention-level is already high as a result of seeing higher variety of novel ads in earlier exposures.

[23] For robustness, we check if the results are significant in a regression with interaction terms and find the same patterns; please see Web Appendix §G.2 for details. We further investigate the source for this heterogeneity and document time-varying user characteristics (e.g., same user when he has seen few vs. many ads) as the main source.

	Dependent variable: Click ($Y^*_{i,t}$)		
	(1)	(2)	(3)
	Past Variety $= V_{i,t-1}$	Past Variety $= \log(V_{i,t-1})$	Past Variety $= \frac{V_{i,t-1}}{t-1}$
Treatment ($W_{i,t}$)	0.00325^{***}	0.00374^{***}	0.00343^{***}
	(7.18)	(5.89)	(7.69)
Past Variety	0.00156^{***}	0.00734^{***}	0.00459^{***}
	(11.31)	(10.83)	(11.66)
Treatment × Past Variety	-0.00033^{**}	-0.00210^{**}	-0.00111^{**}
	(-2.59)	(-2.67)	(-2.93)
Controls in Equation (10)	✓	✓	✓
No. of Obs.	2,405,695	2,405,695	2,405,695
R^2	0.006	0.006	0.006
Adjusted R^2	0.005	0.005	0.005
Note:			*p<0.05; **p<0.01; ***p<0.001

Table 4: Heterogeneity in the effects of variety across past variety.

Implications

We now discuss the managerial relevance and implications of our findings, and the applicability of our framework to other settings.

Managerial Relevance of Variety

Digital platforms now have the ability to deliver numerous interventions to their user base. This automatically gives rise to the importance of variety as a construct since users are exposed to a variety of marketing interventions every day. Additionally, platforms increasingly engage in activities whose by-product is more exploration and increased variety of interventions. These activities include: (1) the use of adaptive exploration/exploitation algorithms to learn consumer taste more efficiently (Lattimore and Szepesvári, 2020), (2) adoption of different fairness criteria to achieve more parity across different demographic groups, which induces more randomization in ad allocation and thereby ad variety (Dwork et al., 2012), (3) employment of algorithms that operate based on increased diversification and randomization in order to prevent polarization (Celis et al., 2019), (4) building algorithms to enhance the reachability of recommender systems, which ensures that the recommender system does not systematically make some of the items out of reach (Dean et al., 2020), and (5) commitment to greater diversity standards in advertising. As these activities increase the variety of marketing interventions, it is more important for managers and platforms to understand the downstream consumer-level consequences of this variety.

Takeaways for Platforms and Advertisers

Our analysis provides insights for both platforms and advertisers. From the platform's perspective, eyeballs or ad impressions are valuable resources and increasing users' attention and clicks leads to higher revenue for them (especially in CPC settings). This explains why major platforms such as Google or Facebook invest heavily in research groups that help build better CTR prediction machines (McMahan et al., 2013; He et al., 2014). At a high level, our paper offers new insights into the CTR prediction problem for platforms by recognizing the causal effects of a new construct of ad variety that helps platform improve their CTR prediction algorithms.

However, the role of ad variety goes beyond just improving CTR prediction machines as it identifies an important externality that has implications for auction design and monetization of ad impressions. Our findings suggest that an increase in ad variety at one exposure changes the likelihood of clicks on future exposures, which violates the assumption of independence of impressions made in commonly used mechanisms such as first- or second-price auctions (see Edelman et al. (2007) and Varian (2007) for well-known examples in the digital advertising context). In light of this externality, it is neither efficient nor optimal to sell an impression to the highest bidding ad, as other competing ads may create greater positive externalities through increasing ad variety. Thus, it is essential for the platforms to develop auctions that incorporate such externalities. A recent paper by Rafieian (2020) presents a revenue optimal dynamic auction mechanism that accounts for the interdependence of impressions and quantifies the loss in platform revenues when this interdependence is ignored.

Our findings also have implications for platforms that use adaptive experimentation tools such as contextual bandits to decide which treatment (i.e., ad copy or promotional content) to show at a given exposure. These approaches often assume that the independence of rewards across treatment arms. However, our findings challenge this assumption and highlight the need to develop more dynamic experimentation approaches (Theocharous et al., 2015).

Finally, while we view this problem through the lens of a platform, our paper also has implications for advertisers. First, our results suggest that advertisers may benefit from showing a variety of creatives, thereby helping advertisers trade-off repeating and varying creatives in their ad campaigns. Second, advertisers can incorporate the information about the effects of ad variety into their decision-making. Although past variety may often be unobserved to advertisers, larger DSPs who bid on behalf of multiple ads can better incorporate our findings. However, all these implications for advertisers have to be interpreted with the necessary caveat that we study click as the main outcome of interest, not conversion.

Attention-based Measures as a Potential Solution

While the discussion above highlights the challenges caused by externalities due to variety effects, our analysis also provides directions for some solutions. In light of our attention-based mechanism, platforms can develop attention-based measures of the form $\lambda_{i,t}$ that are defined at the impression level and capture the past frequency and spacing of ad interventions, i.e., $\lambda_{i,t} = \sum_{s=1}^{t-1} f_t(\mathit{Freq}_{i,s}, \mathit{Space}_{i,t})$.[24] Such attention-based measures can be used in different ways. First, platforms can provide this information as a targeting tool to advertisers. This helps the platform outsource the solution to the externality issue to advertisers and market equilibrium. Second, platforms can use these attention-based measures to approximate the externality that an impression would impose on future impressions and use that to modify the standard first- and second-prices auctions. Third, these measures can be used as contexts in contextual bandits to mitigate the aforementioned issues with these algorithms in light of the effects of ad variety.

Applicability of the Methodological Framework

We now discuss the applicability of our methodological framework to other settings. Within the advertising domain, platforms can use our framework as long as the unconfoundedness of ad allocation is satisfied. While this condition is obviously satisfied in a standard field experiment, a

[24] There are also more direct approaches to measure attention with tracking technologies (McGranaghan et al., 2021).

full experiment is not necessary. It can also be easily satisfied by platforms if they can incorporate a small amount of randomization in their ad allocation mechanism (without significantly hurting their revenues). Many platforms already implement such approaches by adopting ϵ-greedy policies that select the optimal action by $1 - \epsilon$ probability, but gives a non-zero probability to all other actions (Theocharous et al., 2015). An alternative approach is to randomize allocation only for a small portion of their total traffic; some platforms such as Bing use this approach (Ling et al., 2017). Overall, platforms that employ such randomization practices can easily adopt our framework.

Finally, our framework is applicable to domains other than advertising, where we want to study the impact of an increase in variety/diversity. This includes studies at the intersection of digitization and diversity in sequential settings. For example, one could use our framework to study how increased diversity in music consumption affects consumer behavior in music streaming channels, or to examine how app users respond to an increased variety of push notifications. More broadly, our method can help in measuring the effect of treatments that are defined as a function of past behavioral history, i.e., $\textit{Treatment} = f(\textit{Behavioral History})$. These interventions are increasingly relevant as platforms deliver more personalized interventions based on users' past behavior.

Conclusion

In many mobile and digital settings, users are often exposed to a sequence of short-lived marketing interventions within a short period of time. This is particularly true in the context of mobile in-app advertising, where platforms use refreshable ad slots. Users are therefore shown a sequence of potentially different ads within a session and can therefore be exposed to a large variety of ads within the same app-usage session. In this paper, we examine how an increase in the variety of ads shown in a session affects user response to the next ad. We use the quasi-experimental variation in ad assignment in our data and propose a methodological framework that allows us to isolate the effects of an increase in ad variety. We apply our framework to data from the leading in-app ad-network of an Asian country and empirically show that an increase in ad variety increases the CTR on the next ad by approximately 13%, holding everything else constant. We then explore the behavioral mechanism underlying this effect and examine an attention-based account that is based on the past behavioral literature: a novel ad that has been shown less frequently and less recently drives more attention to the advertising slot, thereby generating higher CTR on the next ad. We test various predictions of this behavioral account and find empirical evidence consistent with these predictions. Finally, we discuss the implications of our findings for managers and platforms.

Our paper has certain limitations that present fruitful directions for future research. First, our results establish the effects of variety in the mobile in-app advertising setting. Future research can extend the results to other advertising or non-advertising contexts. Second, we study the problem from a platform perspective. Future research can adopt a more advertiser-focused perspective and examine the role of ad variety from an advertiser's perspective. Finally, the attention based mechanism was postulated after seeing the main effects of variety, and as such the results should be interpreted as confirmatory evidence of one possible mechanism. Our analysis does not rule out other possible mechanisms nor do extensive theory testing by first proposing theories and examining their applicability. Future work may benefit from exploring alternative theories for variety effects.

References

Aravindakshan, A., Peters, K., and Naik, P. A. (2012). Spatiotemporal Allocation of Advertising Budgets. *Journal of Marketing Research*, 49(1):1–14.

Arnosti, N., Beck, M., and Milgrom, P. (2016). Adverse Selection and Auction Design for Internet Display Advertising. *American Economic Review*, 106(10):2852–66.

Braun, M. and Moe, W. W. (2013). Online Display Advertising: Modeling the Effects of Multiple Creatives and Individual Impression Histories. *Marketing Science*, 32(5):753–767.

Bruce, N. I., Murthi, B., and Rao, R. C. (2017). A Dynamic Model for Digital Advertising: The Effects of Creative Format, Message Content, and Targeting on Engagement. *Journal of Marketing Research*, 54(2):202–218.

Bucklin, R. E. and Hoban, P. R. (2017). Marketing Models for Internet Advertising. In *Handbook of Marketing Decision Models*, pages 431–462. Springer.

Celis, L. E., Kapoor, S., Salehi, F., and Vishnoi, N. (2019). Controlling Polarization in Personalization: An Algorithmic Framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 160–169.

Chae, I., Bruno, H. A., and Feinberg, F. M. (2019). Wearout or Weariness? Measuring Potential Negative Consequences of Online Ad Volume and Placement on Website Visits. *Journal of Marketing Research*, 56(1):57–75.

Chen, T. and Guestrin, C. (2016). Xgboost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Choi, H. and Mela, C. F. (2019). Monetizing Online Marketplaces. *Marketing Science*, 38(6):948–972.

Danaher, P. J. (1995). What Happens to Television Ratings During Commercial Breaks? *Journal of Advertising Research*, 35(1):37–37.

Danaher, P. J. and Mullarkey, G. W. (2003). Factors Affecting Online Advertising Recall: A Study of Students. *Journal of Advertising Research*, 43(3):252–267.

Danaher, P. J. and van Heerde, H. J. (2018). Delusion in Attribution: Caveats in Using Attribution for Multimedia Budget Allocation. *Journal of Marketing Research*, 55(5):667–685.

Datta, H., Knox, G., and Bronnenberg, B. J. (2017). Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery. *Marketing Science*, 37(1):5–21.

Dean, S., Rich, S., and Recht, B. (2020). Recommendations and User Agency: The Reachability of Collaboratively-Filtered Information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 436–445.

Deng, Y. and Mela, C. F. (2018). TV Viewing and Advertising Targeting. *Journal of Marketing Research*, 55(1):99–118.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226.

Edelman, B., Ostrovsky, M., and Schwarz, M. (2007). Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords. *American Economic Review*, 97(1):242–259.

eMarketer (2019). Time Spent with Mobile, US.

eMarketer (2020a). Smartphone Users and Penetration in Worldwide.

eMarketer (2020b). US Total Media Ad Spending, by Media.

Godes, D. and Mayzlin, D. (2004). Using Online Conversations to Study Word-Of-Mouth Communication. *Marketing Science*, 23(4):545–560.

Han, S. W. and Marois, R. (2014). Functional Fractionation of the Stimulus-driven Attention Network. *Journal of Neuroscience*, 34(20):6958–6969.

He, X., Pan, J., Jin, O., Xu, T., Liu, B., Xu, T., Shi, Y., Atallah, A., Herbrich, R., Bowers, S., et al. (2014). Practical Lessons from Predicting Clicks on Ads at Facebook. In *Proceedings of the*

Eighth International Workshop on Data Mining for Online Advertising, pages 1–9. ACM.

Helson, H. (1948). Adaptation-Level as a Basis for a Quantitative Theory of Frames of Reference. *Psychological Review*, 55(6):297.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, 71(4):1161–1189.

Hoch, S. J., Bradlow, E. T., and Wansink, B. (1999). The Variety of an Assortment. *Marketing Science*, 18(4):527–546.

Jeziorski, P. and Segal, I. (2015). What Makes them Click: Empirical Analysis of Consumer Demand for Search Advertising. *American Economic Journal: Microeconomics*, 7(3):24–53.

Kahneman, D. (1973). *Attention and Effort*, volume 1063. Citeseer.

Kim, J., Allenby, G. M., and Rossi, P. E. (2002). Modeling Consumer Demand for Variety. *Marketing Science*, 21(3):229–250.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Li, H. and Kannan, P. (2014). Attributing Conversions in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment. *Journal of Marketing Research*, 51(1):40–56.

Ling, X., Deng, W., Gu, C., Zhou, H., Li, C., and Sun, F. (2017). Model Ensemble for Click Prediction in Bing Search Ads. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 689–698.

Little, R. and Yau, L. (1996). Intent-to-treat Analysis for Longitudinal Studies with Drop-outs. *Biometrics*, pages 1324–1333.

McAlister, L. (1982). A Dynamic Attribute Satiation Model of Variety-seeking Behavior. *Journal of Consumer Research*, 9(2):141–150.

McGranaghan, M., Liaukonyte, J., and Wilbur, K. C. (2021). TV Ad Viewability: How Viewer Tuning, Presence and Attention Respond to Ad Content.

McMahan, H. B., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J., Nie, L., Phillips, T., Davydov, E., Golovin, D., et al. (2013). Ad Click Prediction: A View from the Trenches. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Mirrokni, V., Muthukrishnan, S., and Nadav, U. (2010). Quasi-Proportional Mechanisms: Prior-Free Revenue Maximization. In *Latin American Symposium on Theoretical Informatics*, pages 565–576. Springer.

Naik, P. A., Mantrala, M. K., and Sawyer, A. G. (1998). Planning Media Schedules in the Presence of Dynamic Advertising Quality. *Marketing science*, 17(3):214–235.

Nair, H. S., Misra, S., Hornbuckle IV, W. J., Mishra, R., and Acharya, A. (2017). Big Data and Marketing Analytics in Gaming: Combining Empirical Models and Field Experimentation. *Marketing Science*, 36(5):699–725.

Pieters, R., Rosbergen, E., and Wedel, M. (1999). Visual Attention to Repeated Print Advertising: A Test of Scanpath Theory. *Journal of Marketing Research*, pages 424–438.

Rafieian, O. (2020). Revenue-Optimal Dynamic Auctions for Adaptive Ad Sequencing.

Rafieian, O. and Yoganarasimhan, H. (2021). Targeting and Privacy in Mobile Advertising. *Marketing Science*, 40(2):193–218.

Ratner, R. K., Kahn, B. E., and Kahneman, D. (1999). Choosing Less Preferred Experiences for the Sake of Variety. *Journal of consumer research*, 26(1):1–15.

Redden, J. P. (2008). Reducing Satiation: The Role of Categorization Level. *Journal of Consumer Research*, 34(5):624–634.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55.

Rutz, O. J. and Bucklin, R. E. (2011). From Generic to Branded: A Model of Spillover in Paid Search Advertising. *Journal of Marketing Research*, 48(1):87–102.

Sahni, N. S. (2015). Effect of Temporal Spacing Between Advertising Exposures: Evidence from Online Field Experiments. *Quantitative Marketing and Economics*, 13(3):203–247.

Sahni, N. S. (2016). Advertising Spillovers: Evidence from Online Field Experiments and Implications for Returns on Advertising. *Journal of Marketing Research*, 53(4):459–478.

Schumann, D. W., Petty, R. E., and Scott Clemons, D. (1990). Predicting the Effectiveness of Different Strategies of Advertising Variation: A Test of the Repetition-variation Hypotheses. *Journal of Consumer Research*, 17(2):192–202.

Tellis, G. J. (2003). *Effective Advertising: Understanding When, How, and Why Advertising Works*. Sage Publications.

Theocharous, G., Thomas, P. S., and Ghavamzadeh, M. (2015). Personalized Ad Recommendation Systems for Life-time Value Optimization with Guarantees. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Tuchman, A. E., Nair, H. S., and Gardete, P. M. (2018). Television Ad-skipping, Consumption Complementarities and the Consumer Demand for Advertising. *Quantitative Marketing and Economics*, 16(2):111–174.

Tunuguntla, S. and Hoban, P. R. (2021). A Near-Optimal Bidding Strategy for Real-Time Display Advertising Auctions. *Journal of Marketing Research*, 58(1):1–21.

Varian, H. R. (2007). Position Auctions. *International Journal of industrial Organization*, 25(6):1163–1178.

Wilbur, K. C. (2008). A Two-sided Empirical Model of Television Advertising and Viewing Markets. *Marketing science*, 27(3):356–378.

Wilbur, K. C. (2016). Advertising Content and Television Advertising Avoidance. *Journal of Media Economics*, 29(2):51–72.

Wilbur, K. C., Xu, L., and Kempe, D. (2013). Correcting Audience Externalities in Television Advertising. *Marketing Science*, 32(6):892–912.

Yao, S. and Mela, C. F. (2011). A Dynamic Model of Sponsored Search Advertising. *Marketing Science*, 30(3):447–468.

Yoganarasimhan, H. (2013). The Value of Reputation in an Online Freelance Marketplace. *Market-*

ing Science, 32(6):860–891.

Zantedeschi, D., Feit, E. M., and Bradlow, E. T. (2017). Measuring Multichannel Advertising Response. *Management Science*, 63(8):2706–2728.

Web Appendix

A Web Appendix A: Details of Sampling Procedure

We sample a set of sessions from the full data as follows:

- *Users:* To identify users with untruncated history, we split the data into two parts and make two sets of users: (1) $\mathcal{U}_1$ that consists of users generated at least one impression from 30 September 2015 to 21 October 2015, and (2) $\mathcal{U}_2$ that consists of users who generated at least one impression from 22 October 2015 to 30 October 2015. We then sample all users who are available in the second set but not in the first set, i.e., $\mathcal{U}_2 \setminus \mathcal{U}_1$. The fact that the user was not available in the first set suggests that we have the entire observed activity for that user.
- *App:* Within the set of users who satisfy the above condition, we then exclusively focus on their impressions in the most popular app. This is a messaging app that is widely used in the country, and generates over 30% of the total traffic observed in the ad-network. Just focusing on the top app allows us to hold the context of the app constant and cleanly derive the causal effect of the variety of previous ads.

  The empirical CDF of the session length in our sample is shown in Figure A1.

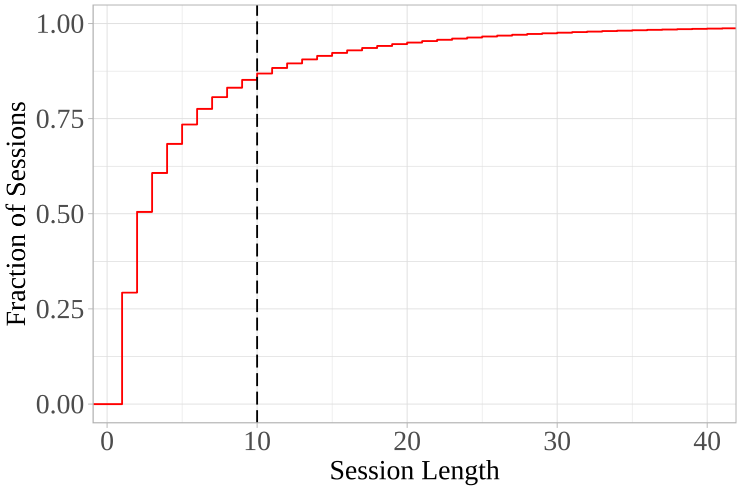


Figure A1: Session length (truncated at 40).

B Web Appendix B: Alternative Specifications for Preliminary Analysis

B.1 Shannon Entropy as an Alternative Measure of Variety

We establish the results presented in Table 2 for a different measure of variety. In particular, we consider Shannon entropy, which is a widely used metric for dispersion and diversity in the information theory literature (Shannon, 1948). In the marketing literature, this measure has been used to quantify dispersion or variety of variables (Godes and Mayzlin, 2004). For a sequence $\langle A_{i,t}\rangle_{t=1}^{T_i}$, let $I_{i,a,t}$ denote the number of times ad a has been shown in session i prior to exposure t. We can define Shannon entropy as follows:

$$\textit{Shannon}_{i,t} = -\sum_{a|I_{i,a,t}>0} \frac{I_{i,a,t}}{t-1} \log\left(\frac{I_{i,a,t}}{t-1}\right). \tag{A1}$$

Intuitively, this measure captures the *amount of information* in the past sequence. More practically, it translates into number of bits required to store the sequence. As such, it takes a higher value when the variety of prior ads is higher. We replace $V_{i,t}$ with $\textit{Shannon}_{i,t}$ and estimate the models in Table 2. We present the results of these models in Table A1. As shown in this table, all qualitative findings in Table 2 remain the same in our new analysis.

	Dependent variable: Click ($Y_{i,t}$)			
	(1)	(2)	(3)	(4)
$\textit{Shannon}_{i,t}$	0.00715***	0.00323***	0.00323***	0.00269***
	(26.35)	(11.22)	(11.20)	(8.48)
$\textit{Freq}_{i,t}$				−0.00035**
				(-2.55)
$\textit{Space}_{i,t}$				0.00035***
				(5.31)
Ad FE	✓	✓	✓	✓
Exposure Number FE	✓	✓	✓	✓
Targeting Variables FE		✓	✓	✓
Session Length FE			✓	✓
R^2	0.007	0.009	0.009	0.009
Adjusted R^2	0.007	0.008	0.009	0.009
No. of Obs.	1,993,542	1,993,542	1,993,542	1,993,542
Note:				*p<0.05; **p<0.01; ***p<0.001

Table A1: Preliminary results on the effect of variety on click outcome when using Shannon entropy to measure variety.

B.2 Preliminary Results with Logistic Regression

Since our outcome variable is binary, we replicate the results of Table 2 with a logistic regression model. We present the estimates of our logistic regression model in Table A2. The results show the same patterns as our main analysis. Please notice that the discrepancy in the number of observations between Table 2 and Table A2 is due to the fact that the logistic regression categories for which the outcome has no variation (e.g., no click for a specific ad).

	Dependent variable: Click ($Y_{i,t}$)			
	(1)	(2)	(3)	(4)
Variety ($V_{i,t}$)	0.15265***	0.06922***	0.06922***	0.05016***
	(0.00569)	(0.00608)	(0.00608)	(0.00642)
$Freq_{i,t}$				-0.05381^{**}
				(0.00914)
$Space_{i,t}$				0.01372***
				(0.00433)
Ad FE	✓	✓	✓	✓
Exposure Number FE	✓	✓	✓	✓
Targeting Variables FE		✓	✓	✓
Session Length FE			✓	✓
No. of Obs.	1,986,048	1,986,048	1,989,602	1,989,602
Note:		*p<0.05; **p<0.01; ***p<0.001		

Table A2: Preliminary results on the effect of breadth of variety on click outcome using a logistic regression model. Robust standard errors are presented in paranthesis.

C Web Appendix C: Proofs

C.1 Proof of Proposition 1

The distribution of propensity scores for ads is fully determined by the allocation rule in a quasi-proportional auction. That is, for any ad a competing for the impression t in session i, the probability that this ad wins an impression is denoted by $\pi_{i,t}(a)$, characterized as follows:

$$\pi_{i,t}(a) = \mathbb{1}(a \in \mathcal{C}_{i,t}) \frac{b_{i,a} m_{i,a}}{\sum_{k \in \mathcal{C}_{i,t}} b_{i,k} m_{i,k}},$$

where $\mathcal{C}_{i,t}$ is the set of ads competing in the auction for exposure t in session i, and $b_{i,a} m_{i,a}$ is ad a's quality-adjusted bid which is the product of the ad's bid ($b_{i,a}$) and quality score ($m_{i,a}$). If we observe all the components of this equation, the distribution of propensity scores for ads is a function of observed variables by default. However, we do not observe the quality scores in our data. So we need to show that the distribution of propensity scores is still fully identified even without observing the quality scores. We first use the following lemma that helps us establish the identifiability of the distribution of propensity scores within a specific sub-sample of our sessions:

Lemma 1. *Let G denote a set of sessions where the auction is identical for any two sessions i and j that $i, j \in G$. The distribution of propensity scores for ads is identified if we observe the actual ad assignment in our data.*

Proof. If we know the actual ad assignments for impressions in G, the proportion of impressions in G that show a in is an accurate estimate of the propensity score for that ad because the distribution of ad assignment is identical across impressions in G. □

In light of this lemma, if we know the actual ad assignments in the data, we do not need to observe bids or quality scores to identify the distribution of propensity scores in any group G where the auction is identical across impressions within that group. Now, if we show that such partitioning or stratification of our sessions is feasible, our proof is complete. We use the fact that the ad-network does not update quality scores throughout the day. Hence, to make sure that the quality scores remain constant in all the impressions within a partition, it is sufficient to partition our sessions such that sessions in different days are not in the same partition. Finally, since we directly observe bids, we only need to show that we can identify $\mathcal{C}_{i,t}$. This set can only vary if sessions are different in two dimensions: (1) targeting characteristics, because some ads may decide to exclude some sessions based on their targeting characteristics, and/or (2) time because some ad campaigns may be unavailable at some points in time. Since we observe all the targeting characteristics as well as the exact timestamp of each impression, we can identify the groups where all the sessions have the same $\mathcal{C}_{i,t}$, and this completes our proof.

C.2 Proof of Proposition 2

We only need to show that if two exposures share the same targeting characteristics and happen the same timestamp τ, their auctions will be identical. Let i and j denote two sessions, and t and t' refer to the corresponding exposure numbers at timestamp τ, respectively, in these two sessions. For the two auctions for these impressions to be identical, we need to satisfy the following condition:

$$\forall a, \;\; \mathbb{1}(a \in \mathcal{C}_{i,t}) \frac{b_{i,a} m_{i,a}}{\sum_{k \in \mathcal{C}_{i,t}} b_{i,k} m_{i,k}} = \mathbb{1}(a \in \mathcal{C}_{j,t'}) \frac{b_{j,a} m_{j,a}}{\sum_{k \in \mathcal{C}_{j,t'}} b_{j,k} m_{j,k}}. \tag{A2}$$

We now show that this equality holds if these two sessions share the same targeting characteristics. This is because all the elements of the LHS and RHS become identical under this condition. We establish this in more detail below:

1. *Equality in bids:* In our setting, ads are only allowed to submit a single bid at any given point in time. Therefore, an advertiser's bid can be different across sessions (i.e., $b_{i,a} \neq b_{j,a}$) if and only if the advertiser changed his bid between the two sessions. Further, if an advertiser changes his bid, it becomes effective only in the next hour. Thus, if i and j are sessions with the same targeting characteristics and impressions (i, t) and (j, t') happen around the same timestamp τ (as broad as an hour), then for all a, $b_{i,a} = b_{j,a}$.

2. *Equality in quality scores:* A unique feature of our setting is that unlike most platforms, this platform does not customize quality scores and only uses an aggregate measure for every ad. Every few days, the platform updates these quality scores. Thus, for sessions i and j in the same day, we have $m_{i,a} = m_{j,a}$. This implies that for two impressions (i, t) and (j, t') with the same targeting characteristics and around the same timestamp τ (as broad as an hour), the we have $m_{i,a} = m_{j,a}$ for all a.

3. *Equality in the set of participants:* If there exists an ad that participates in auction for session i but not session j at timestamp τ (or vice versa), then Equation (A2) is violated. To show that this is not the case for sessions with the same targeting characteristics, we first discuss potential sources of discrepancy in the set of participants and then show that these sources are blocked when i and j share the same targeting characteristics:

   (a) *Difference in targeting:* Advertisers can target their ads based on app category, province, connectivity type, time of the day, MSP, and smartphone brand. As such, each session has a set of targeting characteristics. Hence, we may have $a \notin \mathcal{C}_{i,t}$ and $a \in \mathcal{C}_{j,t'}$ because ad a decided to target session i but not session j (or vice versa). For example, if an advertiser selects Huawei and LG as the set of smartphone brand categories he wants to target, his ad will not be shown to any Samsung users, because Samsung is excluded from her targeting set. Now, if the smartphone brand is Huawei in session i and Samsung in session j, then this ad will be in $\mathcal{C}_{i,t}$ but not $\mathcal{C}_{j,t'}$. However, this source is fully blocked if i and j have the same targeting characteristics. Thus, for any $a \in \mathcal{C}_{i,t}$, we have $a \in \mathcal{C}_{j,t'}$.

   (b) *Difference in availability over time:* An advertiser's campaign availability over time can change due to three possible reasons: entry, exit, and budget. Figure A2 illustrates this point by showing an ad's availability during the time of study. As such, if impression (i, t) occurs during a time when ad a is unavailable (due to entry, exit, or budget) and impression (j, t') occurs during a time when a is available, then we have $a \notin \mathcal{C}_{i,t}$ but

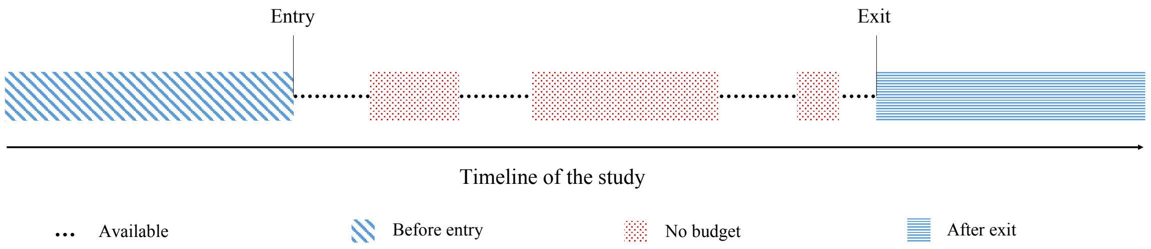


Figure A2: Availability of an ad in the timeline of the study due to entry, exit, and budget exhaustion

$a \in \mathcal{C}_{j,t'}$. However, this source is also blocked if (i, t) and (j, t') happen at the same timestamp τ or in its local neighborhood (empirically, even a gap of an hour does not induce much change in the auctions).

Thus, for impressions (i, t) and (j, t') around the same timestamp with the same targeting characteristics, we have $\mathcal{C}_{i,t} \equiv \mathcal{C}_{j,t'}$.

Together, for impressions (i, t) and (j, t') around the same timestamp with the same targeting characteristics, we have the equality in Equation (A2), and this completes the proof.

D Web Appendix D for Propensity Scores

In this section, we first describe how we estimate the propensity scores for the treatment variable and then show how we assess covariate balance for pre-treatment variables.

D.1 Estimation of Propensity Scores

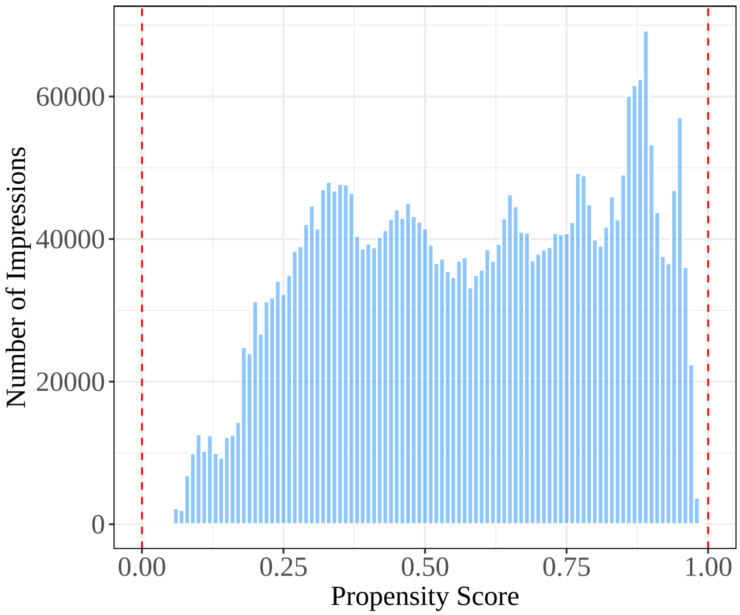


Figure A3: Histogram of estimated propensity scores

We now describe the procedure to estimate the propensity scores of the treatment. As mentioned in the main text of the paper, we can use any learner for this purpose since it is essentially a prediction task. We use XGBoost, which is a fast and scalable version of Gradient Boosting machines that have been used to estimate propensity scores in the past literature (McCaffrey et al., 2013). Like any supervised learning algorithm, XGBoost requires an outcome variable and a set of covariates as inputs for training. The outcome variable or label for this task is the treatment assignment observed in the data ($W_{i,t}$). The set of covariates needed to accurately estimate propensity scores are:

- All the targeting variables $X_{i,t-1}$: province, hour of the day, smartphone brand, MSP, and connectivity type. We include dummy variables for each of these variables. Please notice that the subscript $t-1$ is just for notational consistency, since all these variables largely remain the same for exposure t as well.
- The exact timestamp of the impression to capture any change in the auction, including entry and exit of ads at different points of time.
- The exact latitude and longitude of the user. This is an unnecessary but harmless control.
- A dummy for each ad that indicates whether that particular ad has been shown in the first $t-2$ exposures of the current session. For example, suppose that the sequence of ads shown from exposure 1 to $t-2$ is $\langle A, B, A, C, D \rangle$. Our covariates for all the distinct ads shown A, B, C, and D take value one, whereas for the rest of ads these dummy variables are zero. This set of dummy variables helps incorporate the fact that $e(W_{i,t}) = \sum_{a \notin H_{i,t-2}} \pi_{i,t-1}(a)$.

We use this set of covariates to estimate the propensity scores. To avoid overfitting, we use an early stopping criterion that stops our XGBoost model once two iterations give the same logarithmic loss.

We obtain our propensity estimates and plot their histogram in Figure A3. We see extensive variation in the estimated propensities. Notably, there is no deterministic propensity score, i.e., $0 < \hat{e}(W_{i,t}) < 1$. This ensures that we have the overlap assumption necessary for causal inference.

D.2 Covariate Balance

For covariate balance, we need to ensure that the IPW-adjusted distribution of pre-treatment variables is the same across control and treatment groups. To assess covariate balance, we use the standardized bias measure, which is the commonly used in the literature (McCaffrey et al., 2013). For any pre-treatment variable X, let $\bar{X}_{W=1}$ denote the population mean of variable X when assigned to the treatment. We denote the IPW-adjusted mean of this variable by $\bar{X}_{W=1}^{\textit{IPW-adjusted}}$ and characterize it as:

$$\bar{X}_{W=1}^{\textit{IPW-adjusted}} = \frac{\sum_{j=1}^{N} \frac{\mathbb{1}(W_j=1)}{\hat{e}(W_j)} X_j}{\sum_{j=1}^{N} \frac{\mathbb{1}(W_j=1)}{\hat{e}(W_j)}}, \tag{A3}$$

where j is the subscript for each impression in our data and N denotes the total number of impressions in the sample. Using the definition of $\bar{X}_{W=1}^{\textit{IPW-adjusted}}$, for any variable X, we define the following standardized bias (SB) measure:

$$\textit{SB}(X) = \frac{|\bar{X}_{W=1}^{\textit{IPW-adjusted}} - \bar{X}|}{\sigma_X}, \tag{A4}$$

where $\bar{X}$ is the population mean for variable X, and σ_X denotes its standard deviation. In general, we need to specify a threshold α for the standardized bias such that if $\textit{SB}(X) < \alpha$, we can conclude balance for variable X. The conventional norm in the literature is 0.2 or sometimes 0.1 (McCaffrey et al., 2013; Austin, 2009). We take a more conservative measure and assess balance only if $\textit{SB}(X) < 0.02$.

We check the balance for all our targeting subcategories. Before adjusting for propensity scores, 25 subcategories are unbalanced, i.e., standardized bias is greater than 0.02. As expected, 21 of these subcategories are provinces, and 4 of them are hours of the day. We do not observe any covariate imbalance for subcategories within smartphone brand, MSP, or connectivity type, since these variables were not being used for targeting. After adjusting for propensity scores according to Equation (A3), we achieve balance for all the targeting subcategories.

Next, we check balance for the past variety $V_{i,t-1}$. Again, in the unadjusted case, we are unable to assess covariate balance. However, after adjusting for propensity weights, we assess covariate balance. We further check the balance for the dummy variables indicating whether or not each ad has been shown before in the session. While we have covariate imbalance before adjusting for propensity scores for some ads, we achieve balance for all the ads after adjusting for propensity weights.

E Web Appendix E: Imputation Procedure for Dynamic Selection

We now present the step-by-step imputation algorithm. Let $\tilde{D}$ be the complementary data that we use for imputation.

- *Step 1:* For any session i that ended in T_i exposures, we can impute the timestamp for the exposure $T_i + 1$ that would have happened had the user stayed. We denote this timestamp for the imputed impression as τ^*_{i,T_i+1}. Since each exposure lasts one minute, we add 60 seconds to τ_{i,T_i} to obtain τ^*_{i,T_i+1}.
- *Step 2:* For any timestamp τ^*_{i,T_i+1}, we find an impression j in the complementary data ($j \in \tilde{D}$) with the same targeting characteristics as session i at timestamp τ^*_{i,T_i+1}.[25] Let $\tilde{A}_j(\tau^*_{i,T_i+1})$ denote the ad shown in impression j at timestamp τ^*_{i,T_i+1}.
- *Step 3:* We impute the ad that would have been shown in session i at exposure $T_i + 1$ with the ad found in the complementary data since it represents the ad that could have been shown in the $(T_i + 1)^{\text{th}}$ exposure of session i had it not ended. Hence, $A^*_{i,T_i+1} = \tilde{A}_j(\tau^*_{i,T_i+1}) \sim \mathcal{A}_j(\tau^*_{i,T_i+1})$.

[25] If there are more than one candidate, we randomly choose one impression. Likewise, if there is no impression available at the exact second, we choose the impression that is temporally the closest to τ^*_{i,T_i+1}.

F Web Appendix F: Robustness Checks for the Main Effects

F.1 Robustness Checks with Logistic Regression

Our main results in Table 3 uses a linear probability model. There are several reasons why we focused on linear probability models instead of nonlinear models such as logistic regression. First, we are interested in partial effects for a model where independent variables have a very restricted in the values they can take. In fact, all the covariates are binary variables in our main specifications. Issues with the linear probability models generally arise when independent variables can take extreme values (Wooldridge, 2010). Second, our goal is to estimate partial effects not prediction. As such, we need not worry about the probabilities not lying within the [0,1] range. For the goal of estimating partial effects, we can show that OLS produces consistent and unbiased estimates of coefficients of the linear specification (Wooldridge, 2010). Third, linear probability model flexibly allows for including a very conservative set of covariates and fixed effects while using the full data, which makes it computationally advantageous over nonlinear models such as logistic regression.

However, to check the robustness of our main results, we replicate Table 3 using logistic regression. We present the results of this logistic regression in Table A3. As shown in this table, the results of our logistic regression model generate the same qualitative and even quantitative results.

	Dependent variable: Click ($Y^*_{i,t}$)			
	(1)	(2)	(3)	(4)
Treatment ($W_{i,t}$)	0.11808*** (0.01540)	0.13371*** (0.01349)	0.12839*** (0.01547)	0.14175*** (0.01505)
IPW Adjustment	✓		✓	✓
Imputed Sample	✓	✓		✓
Exposure (t) FE	✓	✓	✓	✓
$Freq_{i,t}$ Indicators	✓	✓	✓	
$Space_{i,t}$ Indicators	✓	✓	✓	
$A^*_{i,t}$ FE	✓	✓	✓	✓
$A_{i,t-1}$ FE	✓	✓	✓	✓
No. of Obs.	2,395,948	2,395,948	1,987,418	2,395,948
Note:		*p<0.05; **p<0.01; ***p<0.001		

Table A3: Average effects of the variety treatment on the CTR using logistic regression. Numbers reported in parentheses are robust standard errors.

F.2 Robustness Checks with Additional Pre- and Post-Treatment Controls

Our main specification takes both pre- and post-treatment confounding into account. Nevertheless, in a series of robustness checks, we now consider even more conservative models with extra controls for pre- and post-treatment variables. One downside of these models is that they ignore much of the exogenous variation in the data and may lack sufficient power. However, they present a conservative test of our main effects – if we are able to find that the effects of variety continue to be directionally correct and significant in these models, it gives us even more confidence in our findings. We now describe these models and present their results in Table A4.

1. **Controlling for targeting variables and past variety:** In the main specification, we control for pre-treatment confounding using IPWs, where the propensity scores are of being assigned to the treatment ($W_{i,t}$) are estimated as a function of all the targeting variables as well as the prior set of ads shown in the session.

   In this model, in addition to the propensity score correction, we also include the covariates that determine the propensity scores directly into the model specification. To do so, we first define *Target* as the interaction of all targeting variables and hour/day. That is, two impressions with the same value of *Target* share the same province, smartphone brand, MSP, connectivity type, and happen at the same hour on the same day. In light of Proposition 2, we know that impressions within a specific targeting area have the same ad allocation distribution. Overall, it gives us 90,074 distinct values for *Target*. Next, as controls for the past set of ads shown in the session, we also include the indicators for different levels of $V_{i,t-1}$. The overall specification of this model is as follows:

   $$\begin{aligned} Y^*_{i,t} =& \beta W_{i,t} + \sum_q \gamma_q \mathbb{1}(\textit{Freq}_{i,t} = q) + \sum_s \delta_s \mathbb{1}(\textit{Space}_{i,t} = s) + \sum_r \theta_r \mathbb{1}(V_{i,t-1} = r) \\ &+ \alpha_0(A^*_{i,t}) + \alpha_1(A_{i,t-1}) + \zeta_t + \eta_{\textit{Target}} + \epsilon_{i,t}, \end{aligned} \tag{A5}$$

   where $\eta_{\textit{Target}}$ captures the fixed effects for *Target*, and θ_r is the coefficient for level r of past variety. We present the results of this model in the first column of Table A4. The significant and positive coefficient for the treatment effects confirms our main effect.

2. **Controlling for user fixed effects and targeting variables:** In this model, we include a different set of controls for the discrepancy in the treatment assignment – user fixed effects and hour/day fixed effects. Controlling for user fixed effects ensures that we only consider the variation within users, so by default it is robust to the selection caused by the differences between users (e.g., users who are more likely to click may be assigned to the treatment condition more often). These fixed effects can also capture the targeting variables that are likely to be fixed within the user (e.g., smartphone brand) and the hour/time of the day (e.g., set of ads targeting that time of day). As before, we also control for the past set of ads shown,

where we use different levels of $V_{i,t-1}$. As such, our model is:

$$\begin{aligned}Y^*_{i,t} =&\beta W_{i,t} + \sum_q \gamma_q \mathbb{1}(\textit{Freq}_{i,t} = q) + \sum_s \delta_s \mathbb{1}(\textit{Space}_{i,t} = s) + \sum_r \theta_r \mathbb{1}(V_{i,t-1} = r) \\ &+ \alpha_0(A^*_{i,t}) + \alpha_1(A_{i,t-1}) + \zeta_t + \eta_{\textit{User}} + \kappa_{\textit{Hour-Day}} + \epsilon_{i,t},\end{aligned} \tag{A6}$$

where $\eta_{\textit{User}}$ and $\kappa_{\textit{Hour-Day}}$ are fixed effects for users and hour-day combination. Overall, this model comes with 71,945 separate user fixed effects and 217 separate hour-day fixed effects. As before, we use an IPW-adjusted regression to estimate this model and present the results in the second column of Table A4. The results show the same pattern as before: an increase in ad variety leads to higher CTR on the next ad.

3. **Controlling for session fixed effects:** Next, we go one step further and make the model even more restrictive by including the session-level fixed effects. It clearly controls for the ad allocation distribution, since the auction is the same for all the seven exposures in a session that we focus on. This gives us 583,694 distinct categories to control for. We use an IPW-adjusted regression to estimate the following model:

$$\begin{aligned}Y^*_{i,t} =&\beta W_{i,t} + \sum_q \gamma_q \mathbb{1}(\textit{Freq}_{i,t} = q) + \sum_s \delta_s \mathbb{1}(\textit{Space}_{i,t} = s) + \sum_r \theta_r \mathbb{1}(V_{i,t-1} = r) \\ &+ \alpha_0(A^*_{i,t}) + \alpha_1(A_{i,t-1}) + \zeta_t + \eta_{\textit{Session}} + \epsilon_{i,t},\end{aligned} \tag{A7}$$

where $\eta_{\textit{Session}}$ controls for session fixed effects. We present the results of this model in the third column of Table A4. Once again, we find that our main effects are robust, even when we use such a narrow lens on our comparison, and a specification that soaks up the exogenous variation across sessions.

4. **Controlling for interaction of all post-treatment variables:** Now, we consider a model with additional post-treatment controls. As discussed before, our approach to address post-treatment confounding is based on a *ceteris paribus* interpretation. Therefore, we now include controls that capture more complex forms of post-treatment confounding. We start by defining a variable $\textit{Current}_{i,t}$ that captures the collective information about the current ad as the interaction of $A^*_{i,t}$, $\textit{Freq}_{i,t}$, and $\textit{Space}_{i,t}$. That is, we make a separate category for each unique combination of these three variables. This gives us 7,174 categories of $\textit{Current}_{i,t}$ that we need to control for. We can write this model as follows:

$$Y^*_{i,t} = \beta W_{i,t} + \alpha_1(A_{i,t-1}) + \zeta_t + \eta_{\textit{Current}} + \epsilon_{i,t}, \tag{A8}$$

where $\eta_{\textit{Current}}$ accounts for the fixed effects of each value of $\textit{Current}_{i,t}$. Note that the above model does not have separate controls for $A^*_{i,t}$, $\textit{Freq}_{i,t}$, and $\textit{Space}_{i,t}$, but instead controls for $\textit{Current}_{i,t}$. We again use an IPW-adjusted regression to estimate our main results. The results for this model are shown in the fourth column of Table A4. Again, we see that the treatment effect is significant and positive, which shows the robustness of our results to more restrictive post-treatment controls.

	Dependent variable: Click ($Y^*_{i,t}$)				
	(1)	(2)	(3)	(4)	(5)
Treatment ($W_{i,t}$)	0.00100*** (4.70)	0.00121*** (5.79)	0.00061** (2.42)	0.00201*** (9.01)	0.00141*** (5.99)
IPW Adjustment	✓	✓	✓	✓	✓
Exposure (t) FE	✓	✓	✓	✓	✓
Freq$_{i,t}$ Indicators	✓	✓	✓		
Space$_{i,t}$ Indicators	✓	✓	✓		
$A^*_{i,t}$ FE	✓	✓	✓		
$A_{i,t-1}$ FE	✓	✓	✓	✓	
Target$_{i,t}$ FE	✓				
$V_{i,t-1}$ Indicators	✓	✓	✓		
User FE		✓			
Hour-Day FE		✓			
Session FE			✓		
Current$_{i,t}$ FE				✓	
Post$_{i,t}$ FE					✓
No. of Obs.	2,405,695	2,405,695	2,405,695	2,405,695	2,405,695
R^2	0.063	0.075	0.257	0.009	0.0571
Adjusted R^2	0.026	0.046	0.018	0.006	0.0015
Note:			*p<0.05; **p<0.01; ***p<0.001		

Table A4: Robustness checks on the average effects of the an increase in ad variety on the CTR on the next ad. Numbers reported in parentheses are t-statistics based on robust standard errors.

Next, we go one step further and define a new variable *Post*$_{i,t}$ by adding another interaction for the previous ad. As such, this variable is the interaction of all four post-treatment controls – $A^*_{i,t}$, $A^*_{i,t}$, *Freq*$_{i,t}$, and *Space*$_{i,t}$. The number of distinct categories in *Post*$_{i,t}$ is 133,845. The model used for this alternative approach is presented below:

$$Y^*_{i,t} = \beta W_{i,t} + \zeta_t + \eta_{Post} + \epsilon_{i,t}, \tag{A9}$$

where η_{Post} controls for all the values of the interaction variable *Post*$_{i,t}$. We use an IPW-adjusted regression to estimate this model and present its results in the fifth column of Table A4. As before, the results show a positive and significant treatment effects.

A few additional points are worth noting here. First, note that all the above models are overly conservative in the sense that they also soak up different sources of exogenous variation depending on their specification. This adversely affects their inferential power. Nevertheless, all of them show a positive and significant effect of variety. Second, in all the models in Table A4, we adjust for IPWs in the regression. If we exclude this adjustment, all the effects become more significant both in terms of the coefficients and t-statistics. Third, we use hourly split of the data to use the results of Proposition 2. It is worth emphasizing that even with shorter splits such as 30 minutes or 10 minutes, the results remain qualitatively the same.

F.3 Robustness Check with Exact Matching of the Sequence of Ads

	Dependent variable: Click ($Y^*_{i,t}$)			
	(1)	(2)	(3)	(4)
Treatment ($W_{i,t}$)	0.00111* (2.49)	0.00118* (2.34)	0.00121* (2.32)	0.00114* (2.19)
IPW Adjustment	✓	✓	✓	✓
Matching Group FE	✓	✓	✓	✓
$A_{i,t-1}$ FE	✓	✓	✓	✓
Freq$_{i,t}$ Indicators		✓		✓
Space$_{i,t}$ Indicators		✓		✓
Targeting Subcategories FE			✓	✓
No. of Obs.	2,405,695	2,405,695	2,405,695	2,405,695
R^2	0.597	0.595	0.597	0.597
Adjusted R^2	0.026	0.026	0.027	0.027
Note:		*p<0.05; **p<0.01; ***p<0.001		

Table A5: Results from exact matching of the sequence of ads. Numbers reported in parentheses are t-statistics computed based on robust standard errors.

In this analysis, we consider an exact matching model to jointly control for both pre- and post-treatment confounding issues, and fully isolate the treatment effect. As such, our goal is to ensure that within units of a matching group, everything is the same except the treatment. To do so, we match the entire sequence of ads except the ad shown at point $t-1$. Let *Matching*$_{i,t}$ denote the matching group that exposure t in session i belongs to. If there is a session j such that *Matching*$_{j,t}$ = *Matching*$_{i,t}$, then we have $\langle A_{i,1}, \ldots, A_{i,t-2}, A_{i,t} \rangle = \langle A_{j,1}, \ldots, A_{j,t-2}, A_{j,t} \rangle$. Hence, the only difference will be in exposure $t-1$, where the treatment assignment happens. This exact matching procedure ensures that all pre-treatment and post-treatment sequence-related factors are identical across within a group, thereby allowing us to fully isolate the treatment effects. Of the 2,405,695 impressions, we are able to match 968,343 of them, with the total of 82,149 separate matching categories.

We control for the *Matching*$_{i,t}$ category for each observation and run a series of different specifications and present the results in Table A5. Since the propensity can still be different despite being in the same matching group, we use IPW-adjusted regression to estimate our treatment effects under this approach. In the first column, we show the results for a model that regresses the click outcome on ad at point t on the treatment, the specific ad shown at the treatment phase and matching group fixed effects. Although there are 1,410,405 matching categories that substantially reduce our statistical power, we still find a significant and positive coefficient for the treatment effects. Next, for the models in columns 2–4, we add more restrictive controls. Even so, the same patterns emerge.

Notice that the total number of categories is higher than the number of categories that we can match, since the former also includes impressions are left unmatched as category with one observation. The categories with just one impression in them do not play a role in our estimation since they are all captured by the matching group fixed effects.

F.4 Alternative Approaches to Dynamic Selection

In this section, we consider alternative approaches to address the dynamic selection problem. As discussed in the main text of the paper, the issue of dynamic selection arises when a user leaves the session right after the assignment to the treatment or control at period $t-1$. The key issue is that the data for these users are missing at the outcome collection phase t. Our main approach in the paper is to impute the specific ad that would have been shown at this missing exposure and assign the outcome zero to it. We now consider alternative approaches to impute the outcome.

	Dependent variable: Click ($Y^*_{i,t}$)		
	(1)	(2)	(3)
Treatment ($W_{i,t}$)	0.00221*** (9.77)	0.00225*** (9.69)	0.00198*** (9.59)
IPW Adjustment	✓	✓	✓
Imputed Sample	✓	✓	✓
Exposure (t) FE	✓	✓	✓
$Freq_{i,t}$ Indicators	✓	✓	✓
$Space_{i,t}$ Indicators	✓	✓	✓
$A^*_{i,t}$ FE	✓	✓	✓
$A_{i,t-1}$ FE	✓	✓	✓
$A^*_{i,t} \times L_{i,t-1}$ FE			✓
No. of Obs.	2,405,695	2,405,695	2,405,695
R^2	0.019	0.005	0.010
Adjusted R^2	0.019	0.005	0.009
Note:	*p<0.05; **p<0.01; ***p<0.001		

Table A6: Average effects of the variety treatment on the CTR using different approaches for dynamic selection and imputation. Numbers reported in parentheses are t-statistics computed based on robust standard errors.

First, we consider a case where we predict what the outcome would have been had the user stayed in the session. This is similar to the practice in biostatistics literature, where the researcher uses the observed data to impute the missing observations (Little and Yau, 1996). As indicated in Challenge 2 in the main text of the paper, we need to impute $\{A^*_{i,t}, Y^*_{i,t}\}$. To do so, we first impute the ad $A^*_{i,t}$ using our approach. We then use a predictive approach to predict the click outcome on this ad had the user stayed in the session. For this purpose, we use an XGBoost classification algorithm, where we give it a large set of inputs and train it on the observed sample. Please see the approach in Rafieian and Yoganarasimhan (2021) for the details of features used in this prediction task. We then take the predictive model to the missing sample and obtain $\hat{Y}^*_{i,t}$ using our predictive algorithm. We impute the outcome ($Y^*_{i,t}$) as the predicted outcome $\hat{Y}^*_{i,t}$ from the XGBoost model. We then estimate the model in the main specification of the paper. The results are presented in the first column of Table A6. The estimated coefficients show the same pattern as before.

Second, we consider another outcome imputation technique based on the users' click decision on the prior ad. A platform may want to distinguish between two types of leave events after the treatment assignment – (1) leave that is caused by the click on the ad shown at period $t-1$, and (2) leave that is not caused by click on the ad shown at period $t-1$. The former is still desirable for the platform, whereas the latter is the event the platform wants to avoid. To reflect this intuition in our outcome imputation, we impute the outcome as one if the user has left the session because of clicking on the ad. In other words, for the user who left at exposure $t-1$, if $Y_{i,t-1}=1$, then we have $Y^*_{i,t}=1$. We give this imputed outcome as the outcome of the main regression and estimate the treatment coefficients. The results shown in the second column of Table A6 reveal the same patterns – an increase in ad variety leads to an increase in CTR on the next ad.

Finally, we consider a case where we distinguish between ad fixed effects in the imputed vs. actual outcomes. This is because the user in the imputed case has not actually seen the ad. We take this account by controlling for the interaction of the current ad ($A^*_{i,t}$) and the leave decision at the prior exposure denoted by $L_{i,t-1}$. We re-estimate our model and present the results in the third column of Table A6. The estimated coefficient shows the same pattern as the main model.

F.5 Clustering Adjustment in Standard Errors

We now discuss the issue of clustering in our estimation of standard errors in the main model. In light of Abadie et al. (2017), we know that a model should adjust for clustering in standard errors if: (1) assignment is clustered, and/or (2) sampling is clustered. Since our sample is the population of interest from a platform perspective, we mainly focus on the former type of clustering. Clustering in assignment means that for two clusters C_1 and C_2 in the data, there is a variance in their treatment assignment propensities. The most notable case is when the randomization is performed at the cluster level. That is, we first decide which clusters receive the treatment (either randomly or through some probabilistic distribution), and then all the users within the same cluster either receive the treatment or control, depending on the cluster assignment. As such, the issue of assignment clustering becomes less relevant if the randomization is performed at the observation level (Jones et al., 2019). This is the case in our empirical setting. Treatment assignment is performed at the impression level, as shown in Figure A3. In fact, the probabilistic allocation rule in our quasi-proportional auction prevents any level of clustering in the treatment assignment because all participating ads have a non-zero propensity of being shown.

However, this does not fully rule out the possibility of clustered treatment assignment. More formally, there is clustering in treatment assignment if the variance of treatment propensity for different clusters is non-zero. Hence, it only suffices to have $\Pr(W = 1 \mid C_j) \neq \Pr(W = 1 \mid C_k)$ for two clusters C_j and C_k in the data. Theoretically, this level of clustering may exist in our study at the level of all the inputs used to estimate the propensity scores in Web Appendix §D.1. However, in our IPW-adjusted regression, we obtain robust standard errors that incorporate all the variation in propensity scores, thereby reflecting clustering in treatment assignment (if any). Thus, we do not need to adjust for clustering in our main model.

Nevertheless, as a robustness check, we consider different plausible scenarios for clustered assignment and obtain clustered standard errors at the level corresponding to each scenario. We do that in addition to the regular robust standard errors in the IPW-adjusted regression model. In light of the discussion above, we expect to see no difference after accounting for different possible clustering in treatment assignment, since all the clustering is already taken into account. To find different plausible clustering in treatment assignment, we use Proposition 2 in the paper, which says that if two impressions share the same targeting characteristics and happen around the same time, their auctions are identical. For the treatment assignment, we know that we need to also take the prior set of ads into account. Thus, we consider the following clustering scenarios:

- Clustering at the interaction of $\mathit{Target}_{i,t}$ and past variety $V_{i,t-1}$ (first column in Table A7).
- Clustering at the interaction of $\mathit{Target}_{i,t}$ (second column in Table A7).
- Clustering at the interaction of Province and Hour-Day, since we know most of the discrepancy and clustering in assignment comes from this (third column in Table A7).
- Clustering at the propensity score level where levels are distinguished by 0.05 margin (fourth column in Table A7).

Together, all the results in Table A7 consistently show that at different levels of clustering, the estimated standard errors do not change. This is what we expected since robust standard errors in our IPW-adjusted regression takes the discrepancy and clustering in treatment assignment into account.

	Dependent variable: Click ($Y^*_{i,t}$)			
	(1)	(2)	(3)	(4)
Treatment ($W_{i,t}$)	0.00186***	0.00186***	0.00186***	0.00186***
	(8.38)	(8.45)	(8.48)	(8.87)
Clustering	*Target*$_{i,t}$ × $V_{i,t-1}$	*Target*$_{i,t}$	*Province* × *Hour* × *Day*	Propensity Score Level (0.05)
IPW Adjustment	✓	✓	✓	✓
Imputed Sample	✓	✓	✓	✓
Exposure (t) FE	✓	✓	✓	✓
Freq$_{i,t}$ Indicators	✓	✓	✓	✓
Space$_{i,t}$ Indicators	✓	✓	✓	✓
$A^*_{i,t}$ FE	✓	✓	✓	✓
$A_{i,t-1}$ FE	✓	✓	✓	✓
No. of Obs.	2,405,695	2,405,695	2,405,695	2,405,695
R^2	0.006	0.006	0.006	0.006
Adjusted R^2	0.005	0.005	0.005	0.005
Note:				*p<0.05; **p<0.01; ***p<0.001

Table A7: Average effects of the variety treatment on the CTR with cluster-robust standards errors. Numbers reported in parentheses are t-statistics computed based on cluster-robust standard errors.

F.6 Placebo Analysis

We now run a placebo analysis with a treatment that most likely has zero impact on the outcome. Our goal is to see if our model returns null effects for a treatment that likely has null effects. To find such a treatment, we use the same treatment variable of an increase in ad variety, but from the previous sessions. That is, if i' is the last session where the user in session i has been at exposure t, we define our placebo treatment to be $\textit{Placebo}_{i,t} = W_{i',t}$. The variable $\textit{Placebo}_{i,t}$ is missing if $W_{i',t}$ does not exist. We replace $W_{i,t}$ in our main specification with $\textit{Placebo}_{i,t}$ and run our model. In addition, we consider other specifications used in robustness check to ensure that those models are also robust to a placebo treatment. We present the results in Table A8. As expected, all the coefficients are null with very small t-statistics. Thus, our placebo analysis provides evidence that our main specification indeed delivers a null estimate for a potentially null effect.

	Dependent variable: Click ($Y^*_{i,t}$)			
	(1)	(2)	(3)	(4)
Placebo ($W_{i',t}$)	0.00004	−0.00000	−0.00003	0.00013
	(0.29)	(−0.03)	(−0.18)	(0.37)
IPW Adjustment	✓	✓	✓	✓
$A_{i,t-1}$ FE	✓	✓		✓
$\textit{Freq}_{i,t}$ Indicators	✓	✓		
$\textit{Space}_{i,t}$ Indicators	✓	✓		
$A_{i,t-1}$ FE	✓	✓		
$\textit{Target}_{i,t}$ FE		✓		
$V_{i,t-1}$ Indicators		✓		
$\textit{Post}_{i,t}$ FE			✓	
Matching Group FE				✓
No. of Obs.	2,000,639	2,000,639	2,000,639	2,000,639
R^2	0.002	0.054	0.054	0.597
Adjusted R^2	0.002	0.013	−0.007	−0.024
Note:		*p<0.05; **p<0.01; ***p<0.001		

Table A8: Average effects of the placebo treatment on the CTR using different approaches. Numbers reported in parentheses are t-statistics computed based on robust standard errors.

G Web Appendix G: Supplementary Results on the Mechanism

G.1 Supplementary Materials for the Section on Treatment Effects Across Different Control Groups

In this section we first present the regression models used to generate Figure 8 in the paper and then show the robustness of those results to alternative specifications.

G.1.1 Regression Tables for Figure 8

We run twelve separate regressions to generate Figure 8 – six separate regressions for Figure 8a and six separate regressions for Figure 8b. In all cases, we use the same treatment group $W_{i,t} = 1$, but different partitions of the control group ($W_{i,t} = 0$) based on their frequency and spacing. In Tables A9 and A10, we show the exact control group that we use to estimate the treatment effects for each column along with the estimates.

	Dependent variable: Click ($Y^*_{i,t}$)					
	Control Condition: *Freq*$_{i,t-1} = k$					
	(1) $k = 1$	(2) $k = 2$	(3) $k = 3$	(4) $k = 4$	(5) $k = 5$	(6) $k = 6$
Treatment ($W_{i,t}$)	0.00128*** (5.58)	0.00247*** (7.02)	0.00383*** (7.34)	0.00628*** (8.97)	0.00699*** (6.66)	0.00604*** (3.49)
IPW Adjustment	✓	✓	✓	✓	✓	✓
Imputed Sample	✓	✓	✓	✓	✓	✓
Exposure (t) FE	✓	✓	✓	✓	✓	✓
Freq$_{i,t}$ Indicators	✓	✓	✓	✓	✓	✓
Space$_{i,t}$ Indicators	✓	✓	✓	✓	✓	✓
$A^*_{i,t}$ FE	✓	✓	✓	✓	✓	✓
$A_{i,t-1}$ FE	✓	✓	✓	✓	✓	✓
No. of Obs.	1,879,546	1,544,301	1,379,397	1,310,135	1,276,831	1,260,508
R^2	0.006	0.006	0.007	0.007	0.007	0.007
Adjusted R^2	0.006	0.006	0.006	0.006	0.006	0.007
Note:				*p<0.05; **p<0.01; ***p<0.001		

Table A9: Average effects of the variety treatment on the CTR when compared to the control group at different levels of past frequency (*Freq*$_{i,t-1}$). Numbers reported in parentheses are t-statistics computed based on robust standard errors.

	Dependent variable: Click ($Y^*_{i,t}$)					
	Control Condition: *Space*$_{i,t-1} = l$					
	(1)	(2)	(3)	(4)	(5)	(6)
	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$	$l = 6$
Treatment ($W_{i,t}$)	0.00278*** (9.60)	0.00184*** (5.93)	0.00212*** (5.46)	0.00107* (2.15)	0.00126* (2.04)	0.00032 (0.36)
IPW Adjustment	✓	✓	✓	✓	✓	✓
Imputed Sample	✓	✓	✓	✓	✓	✓
Exposure (t) FE	✓	✓	✓	✓	✓	✓
Freq$_{i,t}$ Indicators	✓	✓	✓	✓	✓	✓
Space$_{i,t}$ Indicators	✓	✓	✓	✓	✓	✓
$A^*_{i,t}$ FE	✓	✓	✓	✓	✓	✓
$A_{i,t-1}$ FE	✓	✓	✓	✓	✓	✓
No. of Obs.	1,743,404	1,562,283	1,410,639	1,339,617	1,300,769	1,278,123
R^2	0.006	0.007	0.007	0.007	0.007	0.007
Adjusted R^2	0.006	0.006	0.006	0.006	0.006	0.006
Note:				*p<0.05; **p<0.01; ***p<0.001		

Table A10: Average effects of the variety treatment on the CTR when compared to the control group at different levels of spacing (*Space*$_{i,t-1}$). Numbers reported in parentheses are t-statistics computed based on robust standard errors.

G.1.2 Robustness Checks to test Predictions 1 and 2

We now focus on an alternative modeling approach to check the robustness of the results shown in Figure 8. Instead of running separate regressions comparing the treatment with different control groups based on *Freq*$_{i,t-1}$ and *Space*$_{i,t-1}$, we can directly include these variables in the model. However, we must notice that our propensity scores are estimated for the binary treatment $W_{i,t}$. As such, we need to either control for the propensity scores of these new variables or control for the variables determines propensity scores of *Freq*$_{i,t-1}$ and *Space*$_{i,t-1}$. We take the latter approach and control for *Target*$_{i,t}$, since we have shown that two sessions with the same targeting characteristics that happen around the same time have the same ad allocation distribution (Proposition 2). If the ad allocation distribution (i.e., auction) is identical for two sessions, then the assignment to *Freq*$_{i,t-1}$ and *Space*$_{i,t-1}$ is as good as random.

We estimate two models using *Freq*$_{i,t-1}$ and *Space*$_{i,t-1}$ as covariates and present the results in Table A11. Overall, the results shown in this table reveal consistent patterns with our results in Figure 8. The estimated coefficient for *Freq*$_{i,t-1}$ is negative, which is in line with Prediction 1: the more frequently the ad has been shown in the past, the less likely it is for the user to click on the ad shown after that. On the other hand, as predicted by Prediction 2, the coefficient for *Space*$_{i,t-1}$ is positive. The greater the spacing is at point $t-1$, the higher the likelihood of clicking on the ad shown at t.

	Dependent variable: Click ($Y^*_{i,t}$)	
	(1)	(2)
Freq$_{i,t-1}$	-0.00072^{***}	
	(-7.84)	
Space$_{i,t-1}$		0.00025^{***}
		(5.52)
Exposure (t) FE	✓	✓
Freq$_{i,t}$ Indicators	✓	✓
Space$_{i,t}$ Indicators	✓	✓
$A^*_{i,t}$ FE	✓	✓
$A_{i,t-1}$ FE	✓	✓
Target$_{i,t}$ FE	✓	✓
No. of Obs.	2,405,695	2,405,695
R^2	0.056	0.056
Adjusted R^2	0.019	0.019
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table A11: Average effects of *Freq*$_{i,t-1}$ and *Space*$_{i,t-1}$ on the CTR on the next ad. Numbers reported in parentheses are t-statistics computed based on robust standard errors.

G.2 Supplementary Materials for the Section on Heterogeneity Across Past Usage Frequency and Recency

In this section we first present the regression models used to generate Figure 9 in the paper and then show the robustness of those results to alternative specifications, where we use interactions.

We present the results of Figure 9 in Table A12. These are the models used in the main text to generate the results shown in Figure 9. As such, for each column, we focus on a specific sub-sample of the data.

	Dependent variable: Click ($Y^*_{i,t}$)			
	Data Split			
	(1)	(2)	(3)	(4)
	Low Usage Frequency	High Usage Frequency	Low Usage Recency	High Usage Recency
Treatment ($W_{i,t}$)	0.00250***	0.00022	0.00301***	0.00041
	(7.51)	(1.03)	(9.29)	(1.64)
IPW Adjustment	✓	✓	✓	✓
Imputed Sample	✓	✓	✓	✓
Exposure (t) FE	✓	✓	✓	✓
$Freq_{i,t}$ Indicators	✓	✓	✓	✓
$Space_{i,t}$ Indicators	✓	✓	✓	✓
$A^*_{i,t}$ FE	✓	✓	✓	✓
$A_{i,t-1}$ FE	✓	✓	✓	✓
No. of Obs.	1,336,631	1,069,064	1,246,626	1,159,069
R^2	0.009	0.002	0.009	0.003
Adjusted R^2	0.008	0.001	0.008	0.003
Note:				*p<0.05; **p<0.01; ***p<0.001

Table A12: Heterogeneity in the variety effects across usage frequency and recency. Numbers reported in parentheses are t-statistics computed based on robust standard errors.

Next, we consider a regression model where we use the entire sample and identify the heterogeneity in the main effects by estimating the interaction coefficients. To do that, we formally define both variables $UsageFreq_i$ and $UsageGap_i$. As defined in the main text of the paper, $UsageFreq_i$ is the number of impressions the user has seen in the past prior to the current session and $UsageGap_i$ is the time between the last impression in the previous session and the first impression in the current session in terms of hours. Given the skewed distribution of both variables $UsageFreq_i$ and $UsageGap_i$, we take their logs and include them in the regression model with interactions.

We present the results for these alternative specifications in Table A13. The estimated coefficients for interactions reveal the same insights as Figure 9. Most notably, in the fourth column of this model, we notice the same patterns when we control for user fixed effects. Controlling for user fixed effects helps us focus exclusively on within-user variation. That is, our model only uses the variation within each user to estimate the main effects. As shown in the estimates presented in column (4), the results remain the same, suggesting that the source for this heterogeneity is the time-varying user characteristics (e.g., the same user when she has seen fewer vs. more impressions).

	Dependent variable: Click ($Y^*_{i,t}$)			
	(1)	(2)	(3)	(4)
Treatment ($W_{i,t}$)	0.00331***	0.00071**	0.00220**	0.00171*
	(4.70)	(2.71)	(3.09)	(2.35)
$\log(\mathit{UsageFreq}_i)$	−0.00703***		−0.00488***	−0.00740***
	(−67.27)		(−51.33)	(−55.26)
$\log(\mathit{UsageGap}_i)$		0.00766***	0.00409***	0.00431***
		(54.56)	(29.05)	(29.73)
Treatment × $\log(\mathit{UsageFreq}_i)$	−0.00050***		−0.00035**	−0.00045**
	(−3.51)		(−2.75)	(−3.39)
Treatment × $\log(\mathit{UsageGap}_i)$		0.00059**	0.00037*	0.00043*
		(3.04)	(1.99)	(2.22)
IPW Adjustment	✓	✓	✓	✓
Imputed Sample	✓	✓	✓	✓
Exposure (t) FE	✓	✓	✓	✓
$\mathit{Freq}_{i,t}$ Indicators	✓	✓	✓	✓
$\mathit{Space}_{i,t}$ Indicators	✓	✓	✓	✓
$A^*_{i,t}$ FE	✓	✓	✓	✓
$A_{i,t-1}$ FE	✓	✓	✓	✓
User FE				✓
No. of Obs.	2,405,695	2,405,695	2,405,695	2,405,695
R^2	0.019	0.005	0.010	0.086
Adjusted R^2	0.019	0.005	0.009	0.058
Note:			*p<0.05; **p<0.01; ***p<0.001	

Table A13: Heterogeneity in the variety effects across usage frequency and recency using interactions. Numbers reported in parentheses are t-statistics based on robust standard errors.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107.

Godes, D. and Mayzlin, D. (2004). Using Online Conversations to Study Word-Of-Mouth Communication. *Marketing Science*, 23(4):545–560.

Jones, D., Molitor, D., and Reif, J. (2019). What do workplace wellness programs do? evidence from the illinois workplace wellness study. *The Quarterly Journal of Economics*, 134(4):1747–1791.

Little, R. and Yau, L. (1996). Intent-to-treat Analysis for Longitudinal Studies with Drop-outs. *Biometrics*, pages 1324–1333.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, 32(19):3388–3414.

Rafieian, O. and Yoganarasimhan, H. (2021). Targeting and Privacy in Mobile Advertising. *Marketing Science*, 40(2):193–218.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT press.