ESSAYS ON DIGITAL EXPERIENCE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Murat Unal

August 2022

ESSAYS ON DIGITAL EXPERIENCE

Murat Unal, Ph.D.

Cornell University 2022

Ecommerce retailers are increasingly faced with challenges of finding ways to provide a seamless shopping experience to customers. In the first chapter of this dissertation, we focus on the checkout process and study the impact of adopting one-click buying, a feature that reduces the number of steps required to place a purchase order to a single click, on subsequent customer behavior. Using quasi-experimental data over a period of 35 months from an online retailer before and after the launch of one-click buying, we find adopting one-click buying is effective in lifting customer purchases and does so by making treated customers purchase more often as well as more items. The impact of adopting one-click buying on customer purchases post adoption is economically significant, persistent over time, and heterogeneous across customers. Analyzing clickstream data of customer activity online and purchases across product categories, we provide evidence that the increase in purchases is driven by richer engagement through both more visits to the website and more page views upon visit as well as the expansion of purchases across categories. We discuss the implications of our findings for customer experience and targeting.

In the second chapter, we study the impact of online product sampling on customer behavior. Online product sampling offers customers the try-before-you-buy experience to overcome the shortcomings of information loss they experience when purchasing physical experience products such as food, beverages, and apparel in ecommerce. Using quasi-experimental data over a pe-

riod of 13 months from a retailer before and after the launch of online product sampling, we find online product sampling is effective in lifting customer purchases and does so by making treated customers purchase more items per order. We find the impact of online product sampling on purchases post treatment is economically significant, persistent over time, and heterogeneous across customers. Furthermore, we find the impact spills over positively to brand demand and expands to both online and offline channels. We provide evidence that product sampling generates positive affect among treated consumers.

In the third chapter, we study the impact of a policy change by Facebook, a major social media platform, on user behavior. Platform algorithms play an important role in the digital economy. They serve as gatekeepers in social media platforms because they determine visibility, sharing and flow of information. They also affect both intermediaries who publish content to generate traffics to their websites and users who engage to consume and share content on the platforms. Using data from a series of experiments conducted at a publisher's website over a period of 121 weeks before and after the policy change by Facebook on its algorithm, which aimed to improve user engagement, we find the policy change significantly affected user behavior. In particular, after the change went into effect, user engagement decreased significantly insofar as users made significantly fewer clicks on the publisher's website. Using Google search volume data, we provide evidence on the selection mechanism on users through which the policy change by Facebook affected users. We discuss the implications of our findings for platforms, intermediaries and users.

BIOGRAPHICAL SKETCH

Murat Unal was born on March 15, 1984 in Eskisehir. After his primary education he attended the Maltepe Military High School and graduated from it in 2003. He studied Systems Engineering in the Turkish Military Academy and began his career in the Turkish Army as an officer after his graduation in 2007. Between 2007 and 2009 he went through a series of highly specialized training, which includes the Commando Course, the Advanced Commando Course, and the Airborne Instructor Course in the Turkish Army. Furthermore, in 2009, due to his accomplishments, he was selected among all officers to represent the Turkish Army in the elite combat leadership training, the Ranger and Airborne Course, in the US Army. After successfully completing it he started leading Ranger platoons in life-saving counter terrorism missions across various regions in Turkey and abroad. He was promoted to company leader in 2012 and started commanding elite soldiers in preparing, planning, and executing missions under highly stressful and ambiguous conditions. He concluded his career as a Ranger officer in 2014 to pursue his academic interests in the United States.

In August 2014 he started the Industrial Engineering master's program at the Georgia Institute of Technology and graduated from it in May 2015. He earned another master's degree in Management from the Georgia Institute of Technology in August 2018 and started his PhD studies under the guidance of his advisor, Prof. Young-Hoon Park, in Quantitative Marketing at the Johnson School at Cornell University the same month. In fall 2021, he designed and taught a machine learning class to graduate students, which gave him the privilege of sharing his knowledge through teaching. He will join Amazon as an Economist in August 2022.

Bu tezi henüz 14 yaşında bir çocukken askeri lise sınavlarına hazırlanmaya başladığım günlerden itibaren bana evlerini açan, Maltepe Askeri Lisesi'ndeki dört yılım boyunca her hafta sonlarını benimle geçiren, askeri okullardan dereceyle mezun olduğum en gururlu anlarımdan çok sevdiğim subaylığı bırakmaya karar verdiğim en karamsar anlarıma kadar hep yanımda olan, ve en önemlisi bana olan sevgilerini her daim hissettiren, kıymetli halam Zuhal Aydoğdu ve onun değerli eşi, enişte m, Ibrahim Aydoğdu'ya adıyorum. Eminim, sizin desteğiniz olmasaydı, benim hayatım çok farklı bir yönde ilerlemişti ve bugün ben bu tezi yazıyor olamazdım.

ACKNOWLEDGEMENTS

I am indebted to my advisor and committee chair, Prof.Young-Hoon Park for his continuous support, countless hours of reflecting, reading, giving feedback, encouraging, and most of all patience. I am sure I gave him enough reasons to stop working with me, but he continued to support me throughout the entire process. Without his guidance, and patience this thesis would not have been completed. Thank you, Prof.Young-Hoon Park for all your support, and not giving up on me at times when I was struggling to find my direction.

I am grateful to my committee members who were more than generous with their expertise and precious time. Thank you Prof.Nathan Chi-Chung Yang and Prof.Chris Forman for agreeing to serve on my committee and allowing me to benefit from your expertise. Special thanks to Prof.Nathan Chi-Chung Yang, who not only read and offered detailed feedback to every piece of my work, but also graciously utilized his personal network to assist me during my job search.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

FEWER CLICKS, MORE PURCHASES

Since 1997, Amazon has been using its 1-Click ordering technology, which allows customers to purchase online in a single step. The hallmark of the feature is that it removes the most salient friction customers face in the checkout process. Specifically, it allows customers to place orders in a single step without having to re-enter billing, payment, or shipping information at each purchase. Furthermore, it gives customers the option to place orders without adding items to a shopping cart. It has been speculated that this feature alone increased Amazon's sales significantly and the patent was estimated to be valued at $2.4 billion annually (Digiday 2017; Rejoiner 2017).

Amazon's U.S. patent for 1-Click expired in September 2017. Online retailers and platforms have gained the opportunity to incorporate the technology into their businesses without being forced to pay license fees. Indeed, several firms have responded to this opportunity. Magento, for example, adopted one-click buying under "Instant Purchase" and began offering it to its retail clients starting in December 2017 (Magento 2017). More recently, one-click ordering has also made inroads into social media platforms such as Instagram.

Despite its growing importance, several aspects of one-click buying continue to remain a source of debate for ecommerce companies in deciding whether to incorporate this feature into their online stores. Apart from anecdotal estimates, the economic value of offering shoppers the option to order online in a single step is a secret known only by Amazon (Wagner and Jeitschko 2017). To date, no research has examined how adopting one-click buying affects customer behavior. Other related questions also remain unanswered and merit an empirical

investigation. For example, if adopting one-click buying causes changes in purchase amount, does it also cause shifts in other critical metrics, such as number of orders placed and number of items purchased? Is the impact persistent over time? Finally, what type of customers change their behavior most upon adopting the feature?

Beyond establishing the effect of adopting one-click buying on subsequent customer behavior, it is important to understand the underlying mechanisms behind it for several reasons. First, several competing processes could drive customer behavior after adopting one-click buying, which could cause the overall impact to become positive, neutral or negative. For example, the increased convenience through adopting the feature could lead to channel-switching behavior whereby customers would move most of their purchases to online from offline (e.g. Forman, Ghose, and Goldfarb 2009; K. Wang and Goldfarb 2017). Ultimately, this could result in the net impact on purchases to increase, decrease or remain the same. What is more, the ease of placing orders in a single step could lead to higher rates of impulse buying online after adopting one-click buying (e.g. Hui et al. 2013). Higher impulse buying could be accompanied by higher return rates (e.g. Ridgway, Kukar-Kinney, and Monroe 2008), which could have a net negative effect on revenue. Finally, eliminating frictions could lead to a better shopping experience online, which could ultimately increase customers' engagement with the retailer and result in net positive effect on revenue (e.g. Dutta, Jarvenpaa, and Tomak 2003; Parboteeah, Valacich, and Wells 2009). As such, answering which of these potential mechanisms plays a major role in explaining the effect of adopting one-click buying on customer behavior is critical.

The objective of this paper is to seek answers to the above questions. Providing answers to these questions is important because it would generate insights that can guide ecommerce companies in their decisions about introducing one-click buying at their online stores. We address these questions in close collaboration with a retailer in Asia that introduced one-click buying on its website in January 2017.[1] Using quasi-experimental data over a period of 35 months before and after the launch of one-click buying, we apply a two-step identification strategy that combines rolling entry matching (e.g. Witman et al. 2019; Bell, Gallino, and Moreno 2020) with an econometric analysis and estimate the average treatment effect of adopting one-click buying. Furthermore, we obtain individual-level treatment-effect estimates by applying generalized random forests (Athey, Tibshirani, and Wager 2019) and study the heterogeneity of the effects in a data-driven and non-parametric way.

We find adopting one-click buying is effective in lifting customer purchases. The effect on customer purchases is not only economically significant but also persistent over time. On average, customers who adopted one-click buying increased their purchase amount by $86 after 15 months of adoption, compared to a group of control customers. This corresponds to an increase of $5.7 per month. Considering that the average purchase amount per month was about $20 prior to adoption, the effect is equivalent to a 28.5% increase in monthly purchases. Furthermore, we find that adopting the feature increased purchase amount by making treated customers purchase more often and more items. Our findings are robust to potential confounding effects of self-selection and unobservables, different treated and control groups, and different outcomes of purchase behav-

[1] Because Amazon's patent for the 1-click checkout technology was only applicable in the US, our partner firm in Asia was free to introduce the feature on its online channel without any restrictions.

ior.

We also find substantial variation in the treatment effect across customers. Specifically, the magnitude of the increase in purchase amount, 15 months after adoption, ranges from $29 to $158 post treatment. Adopting one-click buying had a limited impact on high-value customers (based on past purchase patterns) but had a larger impact on customers who previously purchased less and visited the online store less often. Moreover, we find customers who adopted the feature within five months of its inception were more responsive to it than those who adopted it afterwards, and the impact was larger on older people.

To uncover the potential mechanisms that might explain our findings, we leverage a variety of data including offline purchases, product returns, customer activity online, and purchases across product categories. Using these rich data and drawing upon the literature on consumer behavior and digital marketing, we explore channel switching, impulse buying, and customer engagement as possible explanations. We find no evidence for channel substitution and impulse buying.

We propose that introducing one-click ordering provides an opportunity for the online retailer to improve customer engagement, which leads to changes in purchase behavior once customers adopt it. We provide evidence consistent with this explanation based on analyzing customer activity online and purchases across product categories. Specifically, after 15 months of adopting the feature, compared to the control group, treated customers visited the website significantly more often, viewed more pages and spent more time upon visit, and deepened their relationship with the firm by expanding their purchases across categories. Jointly, our findings suggest that the improvement in the

checkout process through one-click buying offers customers an enhanced customer experience, which increases their engagement at the ecommerce website and thereby leads to more purchases after adoption.

In the next section we position our paper with respect to extant studies in the literature. In §1.2 we describe our research setting and data. In §1.3 we discuss our empirical methodology. We present our findings and discuss possible explanations for the effect in §1.4 and §1.5, respectively. We offer several robustness checks in §1.6 and conclude in §1.7.

1.1 Literature Review

Our paper is related to several streams of research. Broadly, our study fits into the literature of the impact of technology on consumer behavior and the long-tail literature. Previous research has demonstrated that with the expansion of the Internet, applications of the latest technology can reshape the retail landscape and impact consumer behavior in various ways.

Studies have shown that IT-enabled features and decision aids can significantly influence purchase decision making and satisfaction (e.g., Häubl and Trifts 2000). Moreover, it has been documented that certain features in the online store can urge consumers to make specific type of purchases, such as impulse purchases (e.g., Parboteeah, Valacich, and Wells 2009). De, Hu, and Rahman (2010) demonstrate how consumers' usage of technological features such as search and recommendations tools influence consumer behavior and thereby affect online sales. In their seminal paper, Brynjolfsson, Hu, and Simester (2011) report that consumers' usage of IT-enabled features lead to larger percentage

of sales from niche products, providing evidence for the long-tail phenomenon in ecommerce. More recently, studies have demonstrated how shopping behavior changes with the technological device being used (e.g., R. J.-H. Wang, Malthouse, and Krishnamurthi 2015; Xu et al. 2016; Narang and Shankar 2019) and how introducing artificial intelligence-enabled tools and services can transform consumer behavior in ecommerce (e.g., Gallino and Moreno 2018; Shankar 2018). We add to this growing literature by studying how the one-click checkout technology affects subsequent customer behavior in ecommerce.

Our study also complements and extends the literature on customer experience (e.g., Peter C Verhoef et al. 2009; Lemon and Peter C Verhoef 2016). Research on customer experience can be grouped into two streams. In one stream studies have shown how introducing new marketing strategies and business services can improve the customer experience and thereby influence customer behavior in ecommerce. One such service is the try-before-you-buy experience in the form of free shipping and returns, which has been shown to increase online sales (e.g., Bower and Maxham III 2012) as well as the sales of high-risk products and overall return rates (e.g., Shehu, Papies, and Neslin 2020). Relatedly, Bell, Gallino, and Moreno (2020) report that visiting an apparel retailer's experience-centric offline store, enhances the customer experience and leads to increase in customer spending as well as shopping velocity but decreases the likelihood of returns.

In another stream, studies have shown the impact of technological features on customer experience. Research in this domain has reported how online features, such as decision aids and personalized shopping lists, improve customer experience and thereby lead to increased satisfaction and loyalty to the focal

retailer (e.g., Palmer 2002; Shi and Zhang 2014). One-click buying, as an online feature, arguably has the potential to enhance the customer experience. As such, we contribute to this body of the customer experience literature by studying the causal effect of adopting one-click buying on customer behavior in ecommerce. Furthermore, our research attempts to uncover the potential mechanisms of the effect, which the previous literature has not considered.

Additionally, we contribute to a growing literature on the effects of various marketing interventions on customer behavior using data from quasi-experiments (e.g., Manchanda, Packard, and Pattabhiramaiah 2015; Datta, Knox, and Bronnenberg 2017; Narang and Shankar 2019; Bell, Gallino, and Moreno 2020; Iyengar, Park, and Yu 2022). As identification is especially challenging in settings where the decision to receive the treatment as well as its timing is self-determined, various econometric approaches have been proposed to overcome this challenge. For example, Manchanda, Packard, and Pattabhiramaiah (2015) and Narang and Shankar (2019) assumed a common treatment date for studying the impact of the launch of an online community and for studying the differences between adopters and non-adopters of a retailer's mobile app, respectively. Iyengar, Park, and Yu (2022) also assumed a common treatment date for studying the impact of subscription programs. We contribute to this area of the literature by presenting an alternative approach for solving the identification challenge when double-selection is present. Specifically, instead of assuming a common treatment date, we apply a two-step identification procedure that combines dynamic matching from the first-stage with an econometric analysis on the matched sample in the second-stage. Recently Bell, Gallino, and Moreno (2020) followed a similar approach, whereby they applied risk-set matching in their first-stage.

Finally, a limited number of papers in this domain of the literature have examined heterogeneous treatment effects using machine-learning methods (e.g., Ascarza 2018; Fong et al. 2019; Simester, Timoshenko, and Zoumpoulis 2020; Rafieian and Yoganarasimhan 2021; Iyengar, Park, and Yu 2022). Our paper adds to this stream of research by offering an application that combines machine learning methods with a two-step identification approach to a marketing related problem.

1.2 Research Setting and Data

We obtained the data for our empirical analysis from a retailer in Asia that prefers to remain anonymous. The retailer we collaborated with specializes in certain product categories, such as clothing, personal care products, and shoes, and sells a wide range of consumer goods at both brick-and-mortar and online channels. The retailer launched a single-step checkout technology, which we refer to as one-click buying, on its online store in January 2017.[2]. The introduction of one-click buying was communicated to online customers through mass emails and on the website, and no specific targeting was involved.

The one-click buying feature in our study is the same 1-Click feature that Amazon employs on its website. After joining one-click buying, customers have the option to order products in a single step. Before the launch of one-click buying, checking out at the online store consisted of several steps. Shoppers first had to place items into shopping carts and then had to verify their shipping and billing information separately. Before submitting their order, they also had

[2]To the best of our knowledge, our partner firm was the only business in its category that offered one-click buying at its online store during our study period.

to review and confirm it. One-click buying streamlined this process by bringing down the number of steps required to place an order to one. Importantly, the retailer did not alter the configuration since the feature was introduced and maintained it throughout our data period.

Our data span a period of 35 months, starting from January 2016 to November 2018. It includes a random sample of 977 customers who registered for one-click buying between January 2017 and September 2017 on a voluntary basis without any economic incentives.[3] These 977 customers maintained their registration throughout the data period and constitute the treatment group in our analysis. For the purpose of comparison, we also obtained a random sample of 17,229 customers who had yet to join the feature as of November 2018. The retailer did not target treated customers with different promotions and communications throughout the data period.

The data consist of three parts: transaction data of customer purchase and return behavior, clickstream data of customer activity online, and sociodemographic data. The transaction data contain detailed information about each order made by a customer, that is, when a customer purchased a product and how much she paid for it. The data also include information on products returned and the categories of products purchased and returned. Using transaction data, we define a set of outcome measures associated with customer purchases. Because one-click buying applies to the online channel only, unless specified otherwise, these measures are based on online purchases and are constructed at the customer-month level. Because we are mainly interested in estimating the long-term effects, depending on the analysis, we aggregate these

[3]Because of the non-disclosure agreement we have with the collaborating firm, we are unable to disclose the total number of customers who registered for one-click buying at the firm during the data period.

monthly measures into measures that span multiple months, i.e., 12 and 15 months.

Our primary outcome measure is the amount spent by a customer after adopting one-click buying.[4] We also consider two other measures of customer purchases, that is, number of orders made (order frequency) and number of items purchased, because the change in purchase amount through one-click buying can arise in multiple ways. For example, one-click buying could lift purchase amount due to the increase in order frequency and/or items purchased. Order frequency and items purchased could also change in opposite directions, but the overall change in purchase amount might still be positive.

The second type of data we obtained is online clickstream data. They contain detailed individual-level information on each visit to the website and customer activity upon visit, that is, when a customer visited the website, and which pages (and how long) she viewed on the website. Using these micro-level data, we present a set of measures associated with customer activity online, such as website visits, page views and duration upon visit to explore possible explanations of our findings. However, this data does not include customer activity online during the checkout process, which prevents us from studying the effect of adopting one-click buying on its usage. Finally, our data contain socio-demographic characteristics of customers, for example, age, gender, and address, which we utilize to further control for customer heterogeneity.

[4] All transactions were recorded in the currency of the country in which the headquarters of the company was located. We converted purchase amount to U.S. dollars using the average exchange rate over the data period.

1.3 Empirical Framework

In this section, we first describe our treated and control groups, and then discuss our descriptive analysis, which we follow by an overview of our identification strategy. We next discuss our econometric approach for estimating the average treatment effects and the generalized random forests procedure, which we employ to estimate the heterogeneity of the treatment effects.

1.3.1 Treated and Control Groups

The launch of one-click buying was an exogenous event that happened on a specific date, i.e. January 1, 2017. However, customers had the option to register for the feature at any time post launch. The proportion of adopters over nine months in 2017 are as follows: Jan.: 15.9%, Feb.: 8.1%, Mar.: 5.8%, Apr.: 8.9%, May: 10.8%, Jun.: 8.3%, Jul.: 13.0%, Aug.: 11.3%, and Sep.: 17.9%. Because the time at which a customer was first exposed to one-click buying varies by customer, we do not have common pre-treatment and post-treatment periods but rather define them individually based on the dates treated customers registered for one-click buying. For example, for customers that adopted one-click buying in September 2017, the months before that constitute the pre-treatment period and the time after that is the post-treatment period.

Defining the pre-treatment and post-treatment periods for the control group is more nuanced and requires more care. The first choice is to use the launch date, January 1, 2017, and set the pre- and post-treatment periods accordingly. The downside of this approach is that it causes the control group's baseline and

post-treatment periods to not match the baseline and post-treatment periods of the majority of the treated customers. This is problematic because the two groups are no longer exposed to the same market conditions and attributing the changes in customer behavior to the treatment becomes even more challenging.

We overcome this challenge by applying rolling entry matching. Specifically, for each treated customer, we find a comparison control customer from the entire pool of 17,299 controls that best matches to the treated at the time of her treatment. For example, for customers who adopted one-click buying in September 2017, we find non-adopters with best matching characteristics at that time. Every control customer's pre-treatment and post-treatment periods are then set to her treated counterpart's pre-treatment and post-treatment periods. This way we ensure that every matched pair of treated and control have the same baseline and post-treatment period and there is no imbalance in time periods between the two groups.

1.3.2 Descriptive Analysis

Before we proceed with rolling entry matching, we perform a descriptive analysis by performing a within customer analysis over time. Given that our data tracks customer purchases over a 35 month period, we take monthly purchase measures at the customer level, Y_{it}, as the unit of observation and include an indicator variable, $Treatment_{it}$, that takes value 1 if month t represents the time after one-click adoption for customer i and 0 otherwise. We include fixed effects at the customer level, θ_i, and to account for any seasonal effects and overall market trends we also include calendar-month fixed effects, λ_t. We use ϵ_{it} as the

error term, which we cluster at the customer level. The following two-way fixed effect (TWFE) model describes our analysis:

$$Y_{it} = \beta_0 + \beta_1 Treatment_{it} + \theta_i + \lambda_t + \epsilon_{it}, \tag{1.1}$$

The parameter β_1 captures the average difference in monthly purchase outcomes between treated and non-treated. The results of this analysis are shown in Table 1.1, and suggest that, within customer, joining one-click buying is associated with increases in monthly purchases, order frequency and items purchased by $\$8.312, 0.496, 0.747$, respectively.

Table 1.1: Panel Analysis on Online Purchases

	Purchase Amount ($)	Order Frequency	Items Purchased
	(1)	(2)	(3)
Treatment	8.312***	0.496***	0.747***
	(0.513)	(0.014)	(0.051)
Month fixed effects	Yes	Yes	Yes
Customer fixed effects	Yes	Yes	Yes
No. of customers	18,206	18,206	18,206
Observations	637,210	637,210	637,210

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Robust standard errors are clustered at the customer level and shown in parentheses.

1.3.3 Identification Strategy

Even though these results suggest a strong impact of joining one-click buying, they can not be interpreted as causal findings. The primary challenge we face in identifying the causal impact of adoption is that treatment assignments are not random and customers self-select into treatment. Self-selection can threaten identification if time-constant and time-varying observable and unobservable differences between treated and control groups are correlated with outcome measures and the treatment indicator at the same time. It is possible that this correlation exists to some degree in our context, resulting in potential selection bias. For instance, customers who had more purchases before the launch of one-click checkout might be more likely to join as they might expect to benefit more after registering for one-click buying.

Our identification strategy aims to minimize selection bias by accounting for potential differences between treated and control groups in two steps. In the first step, we employ rolling entry matching (e.g., Witman et al. 2019; Bell, Gallino, and Moreno 2020) to achieve balance on time-independent and time-dependent observable measures between treated and control groups. This procedure results in a sample of matched pairs, who are indistinguishable along observable measures. In the second step, we take our sample of matched pairs and compare each treated customer's post-adoption purchase measures with their control counterpart's measures and obtain treatment effect estimates.

By following this identification strategy, in our main analysis we assume that there are no constant or time-varying unobservables that can simultaneously influence treatment decision and purchase beahvior, which is known in the causal inference literature as the selection on observables or unconfound-

edness assumption (Angrist and Pischke 2008). In §1.6.3 we further investigate the sensitivity of our findings with respect to this assumption by leveraging the panel structure of our data and employing the TWFE approach on our matched sample.

1.3.4 Rolling Entry Matching

We now discuss the first step we take towards identifying the treatment effects of joining one-click buying. The purpose of this step is to address the time-independent and time-dependent observable differences between the two groups and thereby obtain a matched sample of pairs that are indistinguishable along pre-treatment observable measures and characteristics.

Treated customers can adopt one-click buying at different dates in our context. Therefore matched pairs must be formed such that treated and control customers are similar at the time of treatment. Traditional matching methods compute the propensity score, defined as a customer's propensity of adopting one-click buying, at an instance of time and generate matches statically (e.g., Stuart 2010). Instead, we implement rolling entry matching, which allows us to compute the control groups' propensity scores and generate matches dynamically over time. Specifically, for the controls, we take each month t as a potential treatment date and compute their propensity scores at each t separately based on the covariates prior to month t. For the treated, we use their actual treatment month t and compute their propensity score once. The algorithm embedded into rolling entry matching then takes a newly treated customer at time t and matches her to a control customer who is not yet treated at time t and is closest

to her with respect to the propensity score.

This framework allows us to incorporate not only time-constant but also time-varying characteristics into the computation of the propensity score. By computing the propensity score dynamically at various times, the algorithm finds best matches with respect to both time-independent and time-varying characteristics for every treated customer. This way we generate higher quality matches than traditional algorithms, which create matches statically at a specific point in time.

We estimate the propensity score using logistic regression with three sets of covariates. The first set of covariates relates to customer-firm relationship which would be associated with the adoption of a new technology (e.g., Bolton, Lemon, and Peter C Verhoef 2004; Prins and Peter C. Verhoef 2007). We use (1) number of months since having online account (tenure), (2) elapsed time (days) since last purchase (recency), (3) number of purchases made (frequency), (4) purchase amount ($) spent (monetary value), (5) number of items purchased (items), and (6) number of products returned (return) in the 6 month pre-treatment period, computed on a rolling basis. We also include the sum of online activity measures, computed in a rolling manner, in the 6 month pre-treatment period: (7) number of visits to the website (visit), (8) number of pages viewed (page view), and (9) duration of each visit (duration).

The second set consists of a psychographic measure that reflects customer preferences or interests that might affect adopting one-click buying (e.g., Baumgartner 2002). We operationalize this measure to reflect variety seeking and repeat (or replenishment) behavior in customer purchases, because customers might have different interest in one-click buying depending on their purchase

patterns being variety seeking or repetitive. We employ the Shannon diversity index (e.g., Shannon 1948; Boone and Hendriks 2009) and compute its rolling average in the 6 month pre-treatment period to measure the degree of concentration of purchases across product categories: $H = \frac{-\sum_{j=1}^{J} p_j \log(p_j)}{\log(J)}$, where p_j is the proportion of purchases in month t belonging to category $j \in J$.[5] If purchases are concentrated to one category, the index approaches to 0, whereas it takes the value of 1 if purchases are equally divided across categories.

The final set of covariates captures the socio-demographics of customers. We include age, gender, and address for which we use customers' zip codes and classify them into five regions by income per capita in 2016. Because customers' life styles and purchase patterns might differ across regions, these covariates could help control for other unobserved socio-demographics that might affect joining one-click buying, for example, education, income, life style, and so on. Table 2.3 shows the summary statistics of the variables used in rolling entry matching.

Letting $e(x;\beta)$ denote the model for the propensity score parameterized by β, we obtain a sample of matched pairs through the linearized (estimated) propensity score, i.e., log-odds ratio:

$$l(x;\beta) = \ln\Big(\frac{e(x;\beta)}{1 - e(x;\beta)}\Big).$$

This transformation linearizes values on the unit interval and can improve estimation (Imbens and Rubin 2015). Consider customer i who joined one-click buying, i.e., received treatment, at time t. We ask the question whether, for this customer at time t, there is customer i' in the control group such that the dif-

[5]Based on conversations with the retail partner, we decided to classify purchases across five product categories which corresponds the way the ecommerce website was organized and measured the business performance.

Table 1.2: Covariate Means & Balance Before and After Matching

		Before			After		
Variable	Operationalization	T	C	Diff	T	C	Diff
Tenure	Elapsed time (months) since having online account	82.159	56.033	0.543	119.610	106.572	0.086
		(53.446)	(42.126)		(48.948)	(50.497)	
Recency	Elapsed time (days) since last purchase	22.419	8.131	0.900	17.525	18.320	0.042
		(21.393)	(6.774)		(20.230)	(17.634)	
Frequency	Total number of purchase orders 6m. before treatment	4.930	2.148	0.305	5.599	5.323	0.020
		(11.695)	(5.453)		(10.853)	(16.581)	
Monetary Value	Total amount spent ($) 6m. before treatment	116.894	39.896	0.189	119.610	106.572	0.027
		(530.673)	(223.094)		(324.169)	(612.824)	
Items	Total number of items purchased 6m. before treatment	10.502	3.265	0.131	9.924	9.141	0.018
		(76.115)	(16.928)		(41.614)	(44.525)	
Return	Total number of products returned 6m. before treatment	1.160	0.655	0.080	1.315	1.125	0.024
		(7.343)	(5.153)		(7.906)	(7.819)	
Visit	Total number of visits to the website 6m. before treatment	27.111	13.744	0.229	30.504	32.950	0.031
		(68.923)	(45.345)		(69.036)	(88.926)	
Page View	Total number of pages viewed 6m. before treatment	160.921	76.605	0.226	179.546	201.395	0.039
		(453.115)	(272.151)		(456.926)	(652.779)	
Duration	Total duration (min) of website visits 6m. before treatment	186.840	77.178	0.236	208.122	221.703	0.018
		(571.813)	(321.856)		(581.187)	(862.322)	
Diversity	Average of Shannon diversity index 6m. before treatment	0.047	0.021	0.315	0.055	0.054	0.006
		(0.100)	(0.059)		(0.103)	(0.141)	
Age		36.662	34.780	0.187	36.347	36.006	0.034
		(10.644)	(9.410)		(10.493)	(9.641)	
Gender	1 if female, 0 if male	0.819	0.941	0.383	0.860	0.867	0.022
		(0.385)	(0.235)		(0.347)	(0.340)	
Location 1	1 if in Location 1, 0 otherwise	0.428	0.383	0.092	0.414	0.412	0.005
		(0.495)	(0.486)		(0.493)	(0.0.492)	
Location 2	1 if in Location 2, 0 otherwise	0.242	0.247	0.013	0.246	0.260	0.031
		(0.428)	(0.431)		(0.431)	(0.439)	
Location 3	1 if in Location 3, 0 otherwise	0.156	0.174	0.005	0.160	0.149	0.031
		(0.363)	(0.379)		(0.367)	(0.356)	
Location 4	1 if in Location 4, 0 otherwise	0.089	0.096	0.024	0.094	0.091	0.013
		(0.285)	(0.295)		(0.292)	(0.287)	
Location 5	1 if in Location 5, 0 otherwise	0.086	0.100	0.048	0.086	0.089	0.013
		(0.280)	(0.300)		(0.279)	(0.285)	
Observations		977	17,229		806	806	

Note: Observations are at the customer level, and summary statistics refer to the averages of each covariate in the treated and control groups. Standard errors are in parentheses. Diff stands for absolute value of standardized difference in means.

ference (in absolute value) in linearized propensity scores, $l(x_i;\beta) - l(x_{i'};\beta)$, is less than or equal to a threshold u. In our analysis, we perform 1-to-1 matching without replacement and focus on a threshold of $u = 0.05$ (Stuart 2010), meaning that the difference in propensity scores is approximately less than 5%. The matching algorithm achieves a hit rate of 82% and results in 806 unique pairs that are closest to each other in the propensity scores.

The quality of the matching algorithm can be assessed in a few ways. We examine whether the distribution of the propensity scores are similar between the treated and control groups after matching. Figure 1.1 shows the density of the estimated propensity scores by treatment status, before and after matching on the propensity scores. Before matching, the densities share overlap but vary significantly over the range. Matching balances the densities across treatment status to the extent that no bias seems to remain in the difference of the propensity scores between the groups.

We also assess whether the distribution of the covariates are similar across the treated and control groups after matching. Following (Austin 2009) and (Imbens and Rubin 2015), we examine the standardized differences in covariate means between the treated and control groups. Figure 1.2 presents the absolute value of the standardized differences for each variable used for estimating the propensity scores. Matching leads to a substantial improvement in balance. After matching, all of the standardized differences are below 0.1, a degree of balance comparable to what one might expect in a completely randomized experiment (e.g., Stuart 2010; Imbens and Rubin 2015).

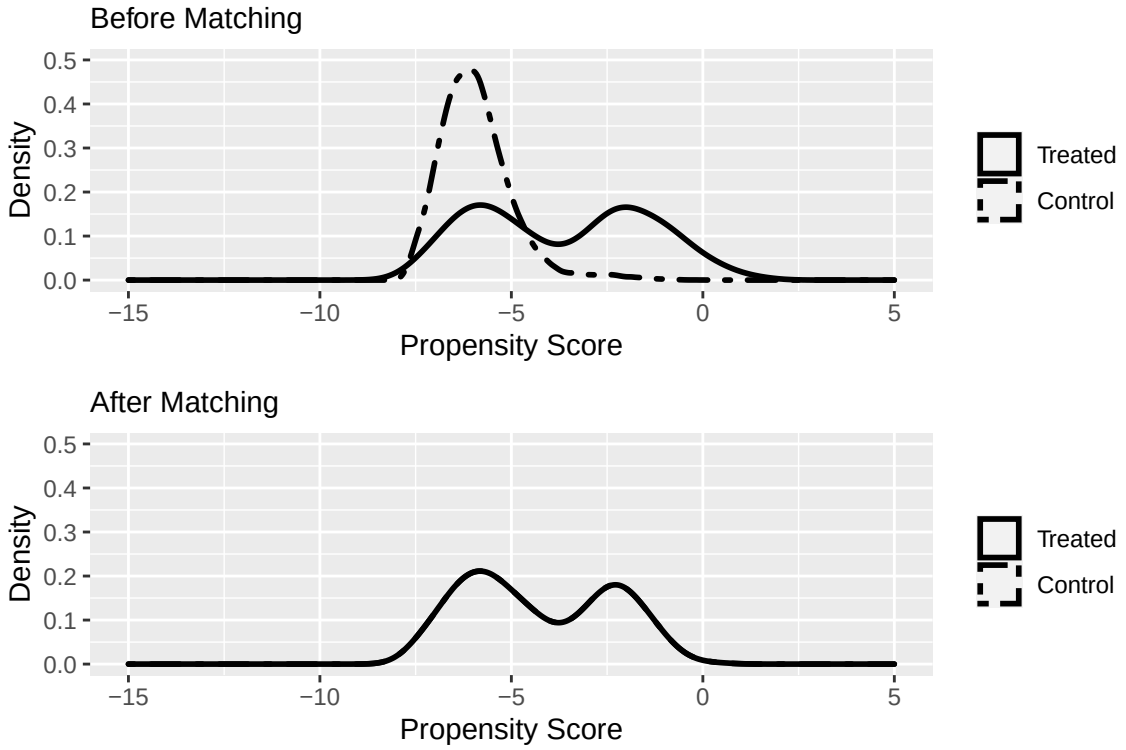

Figure 1.1: Distribution of the Propensity Score

1.3.5 Econometric Analysis

In the second step of our identification strategy we take the matched sample of 806 pairs, and compare the purchase measures of each treated with those of her control counterpart. As described earlier, the post-period is individual-specific and ranges between 15 and 23 months depending on the date a treated customer adopted one-click buying. For January 2017 adopters we have 23 months post-adoption, whereas for September 2017 adopters we only have 15 months post-adoption. Because we have 15 months of post-period observations for every customer, to ease interpretation, we cap the post-period at 15 months and combine monthly purchase measures after treatment into aggregate measures of 12 months and 15 months. This way we are able to study whether the effect

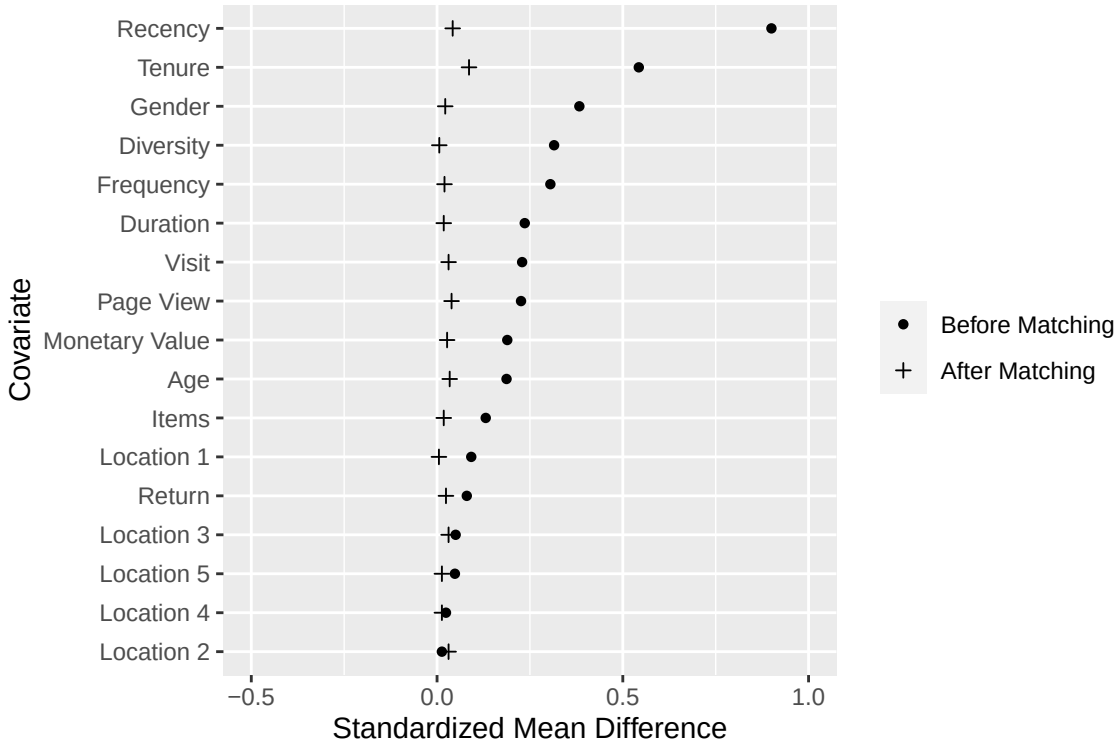

Figure 1.2: Covariate Balance

Note: Absolute values of the standardized mean differences between the treated and control are shown.

of joining one-click buying extends to more than a year after adopting it. The following is our econometric model:

$$Y_{ij} = \beta_0 + \beta_1 Treatment_i + Pair_j + \epsilon_{ij}, \tag{1.2}$$

where Y_{ij} is the purchase measure of customer i belonging to the pair j. $Treatment_i$ indicates whether customer i belongs to the treatment or control group and $Pair_j$ is a pair fixed effect, and ϵ_{ij} is the error term. Notice that in this analysis we compare treated and non-treated within pairs. Since pairs share the same pre and post periods, this way we effectively account for time and market

conditions. We discuss the findings of this analysis in §1.4.1 after we describe our implementation of generalized random forests.

1.3.6 Generalized Random Forests

We are interested not only in estimating the average treatment effects but also in examining their heterogeneity. To this end, we employ a recently developed machine-learning based procedure, called generalized random forests (GRF) (Athey, Tibshirani, and Wager 2019). The procedure is a nonparametric statistical estimation method for causal inference in observational studies and provides a framework to obtain unbiased estimates of average treatment effects and also to capture heterogeneity in parameters of interest. It is an extension of the causal forest method (Wager and Athey 2018), which is based on the classic random forests algorithm used for statistical learning (Breiman 2001).

As a forest-based method for treatment-effect estimation, causal forests apply the same general training and prediction framework used in building random forests, such as resampling, recursive partitioning, and averaging across many trees. What distinguishes GRF from classic random forests is that the splitting criteria for growing individual trees are specifically designed to find partitions where treatment effects most differ. That way, GRF provide a data-driven procedure for selecting the features that are most important for capturing heterogeneity in the treatment effects. Compared to conventional model specifications that rely on interaction terms to capture nonlinear relationships, the method flexibly accommodates complex interplay in data and estimates the parameters of interest without making assumptions on the functional form.

Built on the same general framework as causal forests, GRF critically rely on sample splitting, which is referred to as the honesty condition, and uses different subsamples of the data for growing trees and making predictions at the leaves of the trees. However, the method has an important additional feature that is designed to improve the performance of causal forests. Specifically, instead of obtaining treatment-effect estimates at the tree level for each test example, it creates a list of neighboring training examples and records the frequency by which the test example and each training example share the same leaf in the trees built during training. Based on this information, it assigns (similarity) weights to each neighboring training example and together with their treatment status and outcomes uses them to make predictions for the test example.

A common concern in all supervised machine-learning applications, including causal forests is over-fitting. To prevent over-fitting, we conduct hyper-parameter optimization using cross-validation and perform out-of-bag predictions, meaning for each example, all the trees that did not use this example during training are identified (the example was out-of-bag) and prediction for that example is made using only these trees. For details about GRF, we refer readers to Athey and Wager (2019) and next discuss how we apply it in our context.

We apply the GRF procedure to the aggregated purchase measures described previously. The identifying assumption is the same as in our main analysis, namely the unconfoundedness assumption, but instead of running regressions we apply a non-parametric forest-based procedure to estimate individual treatment effects.

In addition to obtaining individual treatment-effect estimates, another benefit of applying GRF is that the average treatment effects obtained via GRF offer a

robustness check for our main estimates. Because our main econometric model is a linear and additive model, the estimates can be susceptible to violations of this functional form (e.g., Keele 2015). GRF, on the other hand, is a non-parametric procedure that does not rely on any functional form, thus serves as a useful robustness test.

To improve the performance of the causal forest, the developers of the algorithm recommend not to include every available covariate and instead to let the forest search among covariates that are expected to cause heterogeneity in the effect. We follow this recommendation and include a subset of the covariates used in estimating the propensity scores for the heterogeneity analysis. We include tenure, recency, frequency, monetary value, and website visits from the customer-firm relationship set. We also include the diversity index as well as age and gender. Finally, we include the variable entry by dividing the treated into two groups based on their treatment date which allows us to explore how the effect varies by the adoption time of the treated. As half of the treated joined between January and May 2017, we divide the treated into two groups: early adopters that joined in the first five months of 2017 and late adopters that joined between June and September 2017.

1.4 Findings

In this section, we first report our findings for the average treatment effects (ATE) and next discuss the heterogeneity of the effects.

1.4.1 Average Treatment Effects

We report our findings by time period from estimating Equation 1.2 in Table 1.3. Panel A and B show our findings based on 12 months and 15 months post-adoption, respectively. Because our primary interest is to identify the effect of adopting one-click buying for the long term, we discuss the ATE of adoption over 15 months post adoption. Column 1 shows that, treated customers, compared with their matched controls, spent an average of $86 more in the 15 months following their adoption of one-click buying. More granularly, this corresponds to an increase of $5.7 per month. Considering that the average purchase amount per month was about $20 prior to adoption, the effect is equivalent to a 28.5% increase in monthly purchases.

The increase in purchase amount could be driven by the increase in order frequency and/or items purchased. Columns 2 and 3 in Table 1.3 show that treated customers, compared with their matched controls, also placed an average of 6 more orders and purchased an average of 9 more items in the 15 month post-treatment period. On a monthly basis these effects correspond to 0.4 and 0.6 additional orders and items, respectively. Moreover, compared with the pre-treatment monthly values, 0.93 and 1.65, respectively, the effects correspond to 43% and 36% increases in orders placed and items purchased, respectively.

Taken together, our findings of the ATE provide evidence that one-click buying is effective in lifting customer purchases and does so by making treated customers purchase more often as well as more items. The effect on customer purchases is economically significant and persists over 15 months in the post-treatment period.

Table 1.3: ATE on Online Purchases

	Purchase Amount ($)	Order Frequency	Items Purchased
	(1)	(2)	(3)
Panel A. 12 months			
Treatment	62.851*	4.660***	7.315*
	(29.681)	(1.053)	(3.003)
Pair fixed effects	Yes	Yes	Yes
Observations	1,612	1,612	1,612
Panel B. 15 months			
Treatment	86.259**	5.978***	9.146**
	(31.922)	(1.197)	(3.282)
Pair fixed effects	Yes	Yes	Yes
Observations	1,612	1,612	1,612

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Robust standard errors are shown in parentheses.

1.4.2 Heterogeneous Treatment Effects

Using the estimates of the personalized treatment effects obtained through the GRF procedure described in §1.3.6, we next focus on the heterogeneity of the long-term treatment effects.[6] Importantly, the averages of the individual-level treatment effects serve as a robustness check for the estimates presented before, which assume linear and additive treatment effects (e.g., Keele 2015). Table 1.4 reports the ATE as well as their heterogeneity. The ATE are similar to our

[6]Treatment effects in the 12-month post-period have similar levels of heterogeneity.

findings in Table 1.3 from our main analysis.

Table 1.4: Heterogeneity of Treatment Effects on Online Purchases

	Purchase Amount ($)	Order Frequency	Items Purchased
	(1)	(2)	(3)
Mean	96.968**	6.959***	9.240**
	(30.290)	(1.254)	(3.203)
Std.Dev.	37.116	1.812	3.580
Min	29.130	3.066	1.316
Max	157.846	9.439	17.270
N	1,612	1,612	1,612
$N_{\hat{\tau}>=0}$	1,612	1,612	1,612
$N_{\hat{\tau}><0}$	0	0	0

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Standard errors are shown in parentheses.

We find significant variation in the increase in purchase measures across customers as a result of joining one-click buying. Figure 1.3 shows the distribution of the individual-level long-term treatment-effect estimates on purchase amount. The effect is estimated to be positive for all (1,612) customers in our matched sample, but the magnitude of the increase varies considerably, ranging from $29 to $158 post treatment. The effect on both order frequency and items purchased are also positive for all customers and the estimates are similarly heterogeneous. Order frequency increased between 3 and 9 additional orders, and the increase in items purchased varied from 1 to 17 additional items post treatment. These results illustrate the benefits of obtaining individual treatment-

effect estimates by applying the GRF procedure.

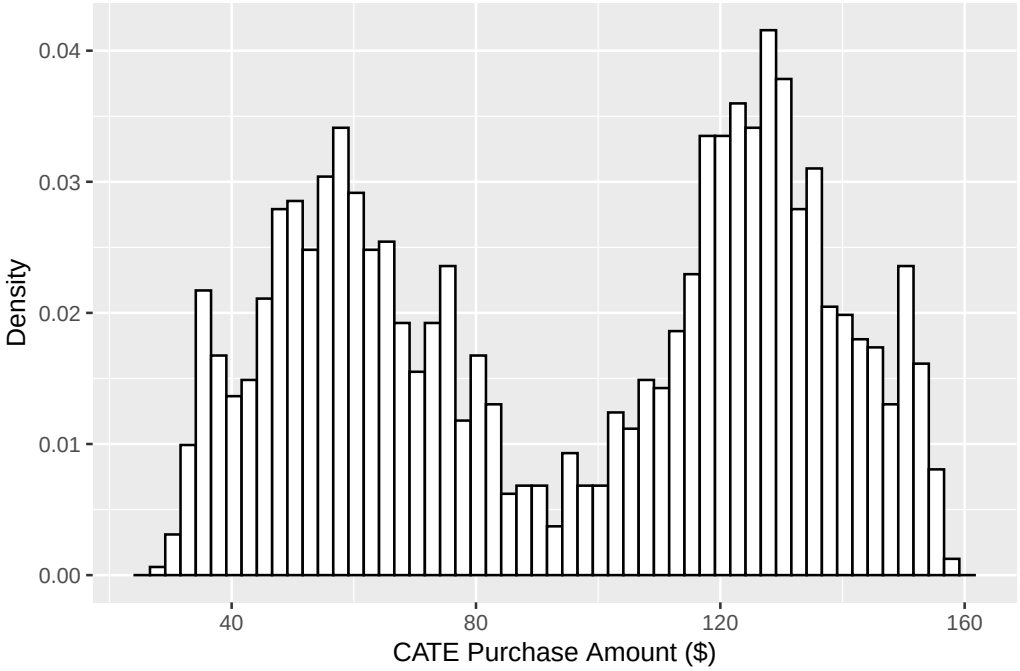

Figure 1.3: Distribution of the Treatment Effect

Since we observe large variation in treatment-effect estimates, we examine the source of the heterogeneity because understanding such variation is important both theoretically and managerially. To that end, we leverage another useful feature that is embedded in GRF, namely the importance measure of covariates used in the estimation. Because GRF build trees during training by splitting on covariates where treatment effects most differ, the importance measure of covariates reflects the relative weight of each covariate in generating the splits. The measure is computed by taking the frequency for which each covariate is used for splitting the nodes and weighting them by the depth of each tree.

Table 1.5 shows the importance weight of each covariate as well as its rank among all covariates used in the analysis. The results suggest that the causal forest spent about 26% of its splits on the entry variable, suggesting that the ef-

fect varies considerably between early and late adopters. More than 50% of the splits were made on the covariates that reflect the customer-firm relationship. In particular, RFM (recency, frequency, monetary value) measures together with website visits, summarize a customer's level of engagement with the firm, and jointly accounted for more than a third of the splits. This result is in line with the findings in the marketing literature that RFM measures could be strong moderators of various marketing activities (e.g., Rossi, McCulloch, and Allenby 1996; Kumar and Shah 2004). Along these lines, tenure was important and accounted for about 18% of the splits. This variable is similar to RFM measures because it reflects customers' level of connection with the online retailer.

Table 1.5: Importance of Covariates in Heterogeneous Treatment Effects

Rank	Importance (%)	Variable
1	26.18	Entry
2	19.38	Age
3	18.04	Tenure
4	12.05	Visit
5	11.61	Recency
6	6.57	Monetary value
7	4.59	Frequency
8	1.57	Diversity
9	0.00	Gender

Interestingly, the causal forest also spent over 19% of its splits on the age variable, suggesting the effect varied between young and old customers. On the

other hand, the diversity index, which measures the extent of variety and repeat behavior in customer purchases across product categories had a limited contribution to the identification of the heterogeneity in our context. This variable reflects customer preferences or interests and is less precise and more nuanced because it is not directly observable (e.g., Baumgartner 2002). Similarly, gender, as another socio-demographic variable, had almost no impact on the splits in our estimation.

Now that we have identified the source of the heterogeneity in the treatment effect, we take a more granular look and seek to identify subgroups of customers for whom one-click buying might have a stronger impact on subsequent purchases. To that end, we relate heterogeneous treatment effects to observed covariates and divide customers into two equal-sized groups based on their covariate values (i.e., below and above the median).[7] We then compute the averages of the personalized treatment-effect estimates across the two groups.

Figure 1.4 shows how the conditional ATE (CATE) differ among the groups with high and low covariate values. Our analysis reveals the biggest difference in the treatment effects exists between early and late adopters, with the former estimated to having a substantially larger effect. Moreover, the figure presents a clear pattern in the changes in purchase behavior with respect to the RFM measures. Customers who purchased more recently, more frequently, and spent more in the pre-treatment period had a smaller increase in purchases after adopting one-click buying, although the differences are statistically not significant. Similarly, customers who made more visits to the online store prior to joining one-click had a smaller treatment effect. In contrast, customers who

[7]For the two binary variables, gender and entry, we divide the observations into their respective classes: male and female; early adopters and late adopters.

had a longer tenure with the firm had a larger treatment effect. In terms of customer-firm relationship, this suggests that joining one-click buying had a stronger effect on those customers who were familiar with the firm but were not necessarily loyalists. Put differently, one-click buying had a limited impact on customers who already had a strong relationship with the firm, but had a larger impact on those customers who previously purchased less in terms of amount and frequency, and visited the online store less often. This is in line with previous work that reports a diminished impact of an experience-centric channel on customers who had more experiences with the firm prior to receiving the treatment (e.g., Bell, Gallino, and Moreno 2020).

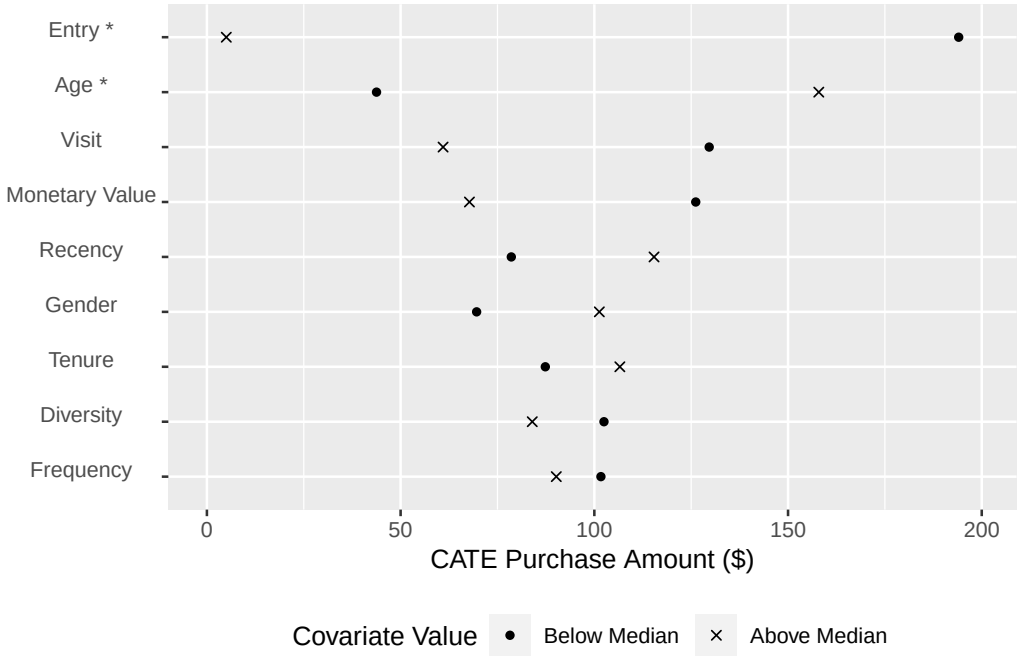

Figure 1.4: ATE for High and Low Values of Covariates

Note: * indicates that the difference of the ATE for high and low values of that covariate is statistically significant at the 5% level. Entry and Gender are binary variables, and we divide observations into their respective classes: male and female; early adopters and late adopters.

With respect to the diversity index, we find the effect is larger among customers who scored lower on variety-seeking behavior. However, this might simply be due to the aforementioned results that customers with low treatment effect had also lower amount and frequency of purchases before treatment. We also find variation in customer demographic characteristics between the high and low treatment effect groups. Specifically, older customers were significantly more responsive to one-click buying than younger customers.

To summarize, we find the impact of one-click buying on customer purchases is heterogeneous across customers. Our findings on heterogeneous treatment effects can assist marketers in scoring and targeting customers and allocating resources across individual customers when designing marketing activity associated with one-click buying.

1.5 Potential Mechanisms

Now that we have established the impact of one-click buying on customer behavior, in this section, we explore some possible explanations underlying the effects.

1.5.1 Channel Switching

Our first explanation is channel-switching behavior. This is based on the possibility that the increase in online purchases through one-click buying was due to the channel-switching behavior to online from offline (e.g., Forman, Ghose, and Goldfarb 2009; K. Wang and Goldfarb 2017). The firm we partnered with is an

omni-channel retailer and is able to link customer purchases between online and offline channels through its reward program. We thus assess channel-switching behavior as a possible explanation underlying the effects.

To that end, we perform the same econometric analysis described in §1.3.5 by replacing online purchases with offline purchases. Columns 1-3 in Table 1.6 present the ATEs on offline purchases. The estimates are statistically insignificant across all outcomes and time periods. These results suggest the increase in online purchases through one-click buying did not come at the cost of offline purchases. Therefore, we conclude that introducing one-click checkout was effective in lifting overall customer purchases for the firm in our context.

Table 1.6: ATE on Offline Purchases and Product Returns

	Purchase Amount ($)	Order Frequency	Items Purchased	Returns
	(1)	(2)	(3)	(4)
Panel A. 12 months				
Treatment	254.682	3.507	14.780	0.063
	(753.250)	(3.052)	(10.617)	(1.380)
Pair fixed effects	Yes	Yes	Yes	Yes
Observations	1,612	1,612	1,612	1,612
Panel B. 15 months				
Treatment	321.288	3.313	14.511	0.030
	(735.496)	(3.012)	(10.406)	(1.386)
Pair fixed effects	Yes	Yes	Yes	Yes
Observations	1,612	1,612	1,612	1,612

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Robust standard errors are shown in parentheses.

1.5.2 Impulse Buying

Another mechanism that might explain the link between one-click buying and increase in purchase outcomes is impulse buying, a purchase that is unplanned, the result of an exposure to a stimulus and decided on the spot (e.g., Piron 1991). In his seminal work, Stern (1962) argues that impulse buying is related to ease of buying to the extent that a purchase involves less of one's resources, such as money, time, and effort. Impulse buying can be triggered by a variety of stimuli ranging from the product itself and its attributes to the environment in which the purchasing event takes place (e.g., Hui et al. 2013). Studying impulse buying in online settings, Parboteeah, Valacich, and Wells (2009) find that a web interface with high-quality task-relevant characteristics, which make shopping efficient and effective, increases the likelihood of customers to buy impulsively. Given that one-click buying adds to an online store's task-relevant features, it might lead to increased impulse purchases.

An additional way in which one-click buying could lead to impulse buying is by hampering customers' mental accounting (Thaler 1985). Customers exert more time and effort for the act of buying when multiple steps are involved in finalizing a purchase. While making buying more burdensome, multiple-step ordering as opposed to one-click buying could provide ample opportunities for customers to deliberate their choice and register the cost of the transaction in their mind. However, when the ordering process is reduced to a single step, the convenience of shopping can override the cognitive processes that keep track of spending, and the losses involved in the purchase might not be apparent, thereby increasing impulse buying. Dutta, Jarvenpaa, and Tomak (2003) show the number of steps involved in the payment process has a significant impact on

the subjects' recall of past expenses. Compared to subjects in the multiple-step payment condition, those under the one-step payment condition have lower recall of past expenses, which in turn leads to impulse buying.

These findings suggest that, impulse buying could contribute to the increase in purchases after adopting one-click buying. To evaluate this explanation, we need a measure of impulse buying. Because our data lack a direct measure of impulse buying, we utilize data on product returns at the online channel and use product returns as a proxy for impulse purchases. The reason we use product returns as a proxy is because it could be positively associated with impulse buying. In fact, Ridgway, Kukar-Kinney, and Monroe (2008) study impulse buying as a dimension of compulsive buying and report that it causes behaviors such as hiding and making frequent returns of purchased items due to feelings of remorse or guilt.

To test for impulse buying, we repeat the econometric analysis from §1.3.5, whereby we use the number of product returns as the outcome variable. Column 4 in Table 1.6 shows the results. Similar to the results for offline purchases, we find that one-click buying had no effect on return behavior for treated customers.[8] As such, these results indicate a lack of evidence that the increase in purchases through one-click buying can be attributed to impulse purchases in our context.

Aside from informing us about the lack of impulse buying, learning that adopting one-click buying does not increase returns offers another important insight. Specifically, the finding makes it unambiguous that the overall impact

[8]We constructed another measure of impulse buying based on the proportion of products returned among products purchased, and did not find any difference in this measure. We also analyzed return behavior at the category level, and did not find any evidence for differences in return behavior at the category level as well.

of adopting the feature is net positive for the firm, which could be the ultimate decision criteria for online retailers in the debate whether introducing one-click buying is worthwhile.

1.5.3 Customer Engagement

As another potential explanation, we propose to examine the impact of one-click buying on customer experience that materializes as customers interact with the retailer. Each of the customer-retailer interactions could evoke customer responses across multiple dimensions (e.g., cognitive, emotional), which then influence customer satisfaction and engagement (e.g., Peter C Verhoef et al. 2009).

Among the three stages in the customer's journey with the firm, namely, pre-purchase, purchase, and post-purchase, the purchase stage is key in shaping customer experience because it is related to a variety of behaviors, such as choice, ordering, and payment (e.g., Lemon and Peter C Verhoef 2016). Online retailers, equipped with useful features such as online decision aids, can facilitate and enable customers to attain their shopping goals.[9] Customers exert less effort and yet make better decisions when online decision aids are available (Häubl and Trifts 2000). Moreover, usage of online decision aids is associated with increased sales (De, Hu, and Rahman 2010) as well as higher satisfaction and likelihood of repeat visits to the website (e.g., Palmer 2002; Parboteeah,

[9]One example of a decision aid is to offer customers the ability to sort alternatives or filter them with certain criteria (e.g., price, brand). Another example involves the option to create personalized shopping lists, or to retrieve previous order lists and place orders using them. Similarly, online retailers often employ recommendation agents that pre-select a set of products that are likely to be attractive for a given individual. Marketing researchers have documented that such interactive decision aids online can influence consumers' information search processes, purchase outcomes, and satisfaction (e.g., Häubl and Trifts 2000; Shi and Zhang 2014).

Valacich, and Wells 2009). Similar to online decision aids, one-click buying is an online feature that offers value to customers by reducing the time and effort during checkout and improving the usability of an online store (e.g., Dutta, Jarvenpaa, and Tomak 2003; Parboteeah, Valacich, and Wells 2009). This increased convenience could lead to a better shopping experience online, which would ultimately increase customers' engagement with the retailer.

One-click buying can drive customers' engagement with the focal retailer in several ways. First, treated customers could become more engaged simply due to the improved shopping experience that the feature offers. Second, the improvement at the online store could lead to a reactivation in online purchase and activity among some of the treated who had been less active for some time period. In this case, the treatment effects could be expected to be stronger among those customers who were familiar with the firm but scored relatively low on the RFM measures prior to their treatment. Finally, if customers had been shopping regularly at other retailers, the convenience through one-click buying at the focal firm might have motivated some customers to switch from other retailers and become more engaged at the focal firm's online store. [10]

Richer engagement could manifest itself through customers' tendency to make more visits to the retailer's website and spend more time upon visit. Providing evidence for this proposed explanation of customer engagement requires clickstream data of customer activity for all customers in our study over the same period of 35 months (between January 2016 and November 2018), which we obtain and complement with purchase data. Using these micro level data, we construct three measures of online activity, that is, visit instances to the web-

[10]Because our partner firm was the only firm in the given product category to offer one-click buying during our study period, this is a probable explanation. However, we do not have access to data from other retailers and cannot confirm this explanation.

site, number of page views, and duration of visits on the retailer's website. In order to ensure that these constructs reflect customer engagement and are distinct from online purchases, we excluded customer activities that involved checkout and payment. In addition, we construct a measure of category expansion by computing the number of product categories a customer made purchases from, because customer engagement could also be reflected in customers' tendency to expand their purchases across product categories.

To test our suggested explanation, we employ the econometric analysis from §1.3.5 as before. Table 1.7 shows the results. We find that, treated customers, compared with their control counterparts, on average made 32 more visits to the online store after joining one-click buying in the long term. They also viewed on average 250 more pages and spent on average 245 minutes more upon visit. We also find treated customers, purchased on average from 3 additional categories, suggesting that they deepened their relationship with the firm by expanding their purchases across categories in the long term.

Our findings are in line with earlier findings that customers become more engaged with the focal firm as their interactions with purchase touch points (e.g., decision aids) improve (e.g., Shi and Zhang 2014). The enhanced shopping experience leads customers to not only increase their purchases but also expand the categories they purchase from. Expanding categories naturally involves more choices, which calls for more time and engagement at the online store. Furthermore, combining these findings with our findings from the heterogeneity analysis in §1.4.2, which showed larger effects among customers who had a longer tenure with the firm but scored lower on the RFM measures, supports customer engagement as being the main driver of the treatment effects.

Table 1.7: ATE on Online Activity and Category Expansion

	Visit Sessions	Page Views	Duration	No. of Categories
	(1)	(2)	(3)	(4)
Panel A. 12 months				
Treatment	27.146***	213.253***	201.097***	2.25***
	(6.788)	(41.410)	(59.928)	(0.341)
Pair fixed effects	Yes	Yes	Yes	Yes
Observations	1,612	1,612	1,612	1,612
Panel B. 15 months				
Treatment	32.232***	250.195***	244.968***	2.888***
	(7.275)	(45.035)	(62.320)	(0.373)
Pair fixed effects	Yes	Yes	Yes	Yes
Observations	1,612	1,612	1,612	1,612

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Robust standard errors are shown in parentheses.

1.6 Robustness Checks

In this section we analyze the robustness of our findings. First, we evaluate the sensitivity of our results with respect to the matching parameter. Second, we analyze whether modifying the start of the post-period changes our findings. Third, we test to what extent our findings are robust to an alternate identification approach. Finally, we reiterate that our results are similarly valid under non-parametric estimation.

1.6.1 Alternate Matching Parameter

The selection of the caliper parameter, u, plays an important function in rolling entry matching, mainly because it decides the range of propensity scores to be considered for matching. A value close to 0 provides stricter requirements for matching. While this is useful for reducing potential biases, it increases the likelihood that some treatment observations will not be matched to a control. Because our priority is to minimize bias, we set u to 0.05 in our main analysis, which resulted in 806 unique matched pairs. As a robustness check, we increase the flexibility of the matching algorithm by changing u to 0.10. This way we obtain a slightly larger matched sample that consists of 811 unique pairs, and run the analysis on this sample. The results are shown in Table 1.8, and are similar to our main findings.

1.6.2 Alternate Outcomes

As another robustness check, we modify the operationalization of the outcomes and perform our analysis by excluding the month in which customers joined one-click buying. If customers adopted one-click buying during the checkout process, they would have additional purchases even if their purchase behavior did not change post adoption. We thus investigate to what extent our findings are sensitive to this possibility. The results are shown in Table 1.9 and are similar to our main findings. This suggest that the treatment effects are not artifacts of the purchases made during the adoption stage, but rather reflect the changes in subsequent customer behavior.

Table 1.8: ATE Alternative Matching

	Purchase Amount ($)	Order Frequency	Items Purchased
	(1)	(2)	(3)
Panel A. 12 months			
Treatment	63.010*	4.694***	7.345*
	(29.35)	(1.039)	(2.987)
Pair fixed effects	Yes	Yes	Yes
Observations	1,622	1,622	1,622
Panel B. 15 months			
Treatment	85.179**	5.942***	9.100**
	(31.236)	(1.166)	(2.987)
Pair fixed effects	Yes	Yes	Yes
Observations	1,622	1,622	1,622

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Robust standard errors are shown in parentheses.

1.6.3 Alternate Identification

Our two-step identification strategy hinges on the unconfoundedness assumption, which states that controlling for observable differences between the treated and control sufficiently accounts for any dependencies between the decision to adopt one-click buying and potential outcomes. This assumption fails, however, if there are unobservable factors that influence the adoption decision and outcomes simultaneosuly, and we are unable to control for them in our analysis.

Table 1.9: ATE Excluding First Month Post Treatment

	Purchase Amount ($)	Order Frequency	Items Purchased
	(1)	(2)	(3)
Panel A. 12 months			
Treatment	59.064*	4.396***	6.924*
	(24.688)	(1.014)	(2.880)
Pair fixed effects	Yes	Yes	Yes
Observations	1,612	1,612	1,612
Panel B. 15 months			
Treatment	82.472**	5.712***	8.756**
	(26.382)	(1.111)	(3.090)
Pair fixed effects	Yes	Yes	Yes
Observations	1,612	1,612	1,612

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Robust standard errors are shown in parentheses.

To evaluate the robustness of our findings with respect to this possibility, we utilize the panel structure of our data and implement the same within customer analysis from §1.3.2. This approach allows us to to account for each individual's characteristics that are unobservable to us and are constant over time, but might simultaneously affect outcomes and adoption decisions. Different from the analysis in §1.3.2, this time we implement the two-way fixed effect (TWFE) model in the second step of our identification strategy, using the panel data of our matched sample of 806 pairs instead of using our unmatched sample.

The results of this analysis are presented in Table 1.10, and show that within customer, joining one-click buying increases monthly purchases, order frequency and items purchased by $\$4.577, 0.336, 0.462$, respectively.

Table 1.10: ATE Using TWFE

	Purchase Amount ($)	Order Frequency	Items Purchased
	(1)	(2)	(3)
Treatment	4.577*	0.336***	0.462*
	(1.946)	(0.076)	(0.193)
Month fixed effects	Yes	Yes	Yes
Customer fixed effects	Yes	Yes	Yes
No. of customers	1,612	1,612	1,612
Observations	56,420	56,420	56,420

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Robust standard errors are clustered at the customer level and shown in parentheses.

By adopting this identification strategy we assume that in the absence of treatment, the expected changes in outcome measures of the treated would be the same as the expected changes of the control. Known as the parallel-trends assumption in the literature on causal inference , it is the foundation upon which the TWFE model is built on and enables us to identify the effect of adopting one-click buying on customer behavior among treated customers in an alternative way. Even though the parallel-trends assumption is impossible to verify, assessing its validity can be done by comparing the trends of the treated and control groups in the pre-treatment period (Angrist and Pischke 2008). Since

the assumption will fail if time-varying unobservables affect the outcomes of the treated and control differently, the pre-treatment period will be useful in evaluating this possibility. If trends are similar in the pre-treatment period, it is likely that they would have followed similar paths in the counterfactual post period without treatment, but joining one-click has resulted in a deviation from the common trend for the treated customers in the factual post-treatment period.

To evaluate whether the parallel-trends assumption is credible in our context, we estimate the treatment effect of one-click buying on purchases in the pre-treatment period by updating Equation 1.1 with the following model:

$$Y_{it} = \theta_i + \lambda_t + \sum_{j=12}^{1} \beta_j Preperiod_{ijt} + \epsilon_{it}, \tag{1.3}$$

where $Preperiod_{ijt}$ takes the value of 1 if customer i joined one-click buying j periods after t, and 0 otherwise. We normalize twelve periods before treatment ($j = 12$) as the baseline of 0 and estimate treatment effects over the remaining 11-month pre-period. Our primary interest in this estimation are the β_j parameters, which capture the average difference in purchase outcomes between treated and non-treated in $Preperiod_j$ relative to the baseline. To the extent that purchase trends are common prior to one-click buying, estimates of β_j should result in a null effect in all preperiods. Similar to Equation 1.1 we include individual fixed effects as well as period fixed effects. We also use robust standard errors clustered at the customer level to account for any serial correlation (e.g., Bertrand, Duflo, and Mullainathan 2004).

The results of this regression are shown in Table 2.4. There is strong support

for parallel trends, 32 out of the 33 coefficient estimates in the pre-treatment period are statistically not different from zero at the 5% level.

1.6.4 Alternate Functional Form

Another robustness test involves evaluating the linear and additive structure of our main estimation. The estimation via GRF allows us to relax this assumption, because of its non-parametric structure, and as we discussed in §1.4.2, our findings are not influenced by this assumption.

1.7 Conclusions

The end of Amazon's hold on the one-click checkout technology offers opportunities to retailers, and naturally leads to the question about the economic value of introducing this technology. Through our partnership with an omnichannel retailer that launched one-click buying on its website, we utilize quasi-experimental data over a period of 35 months and measure the causal effect of adopting one-click buying on customer behavior post adoption. Our two-step identification strategy leverages the longitudinal aspect of our data and is based on first creating matched pairs of treated and controls and then estimating the effects on the sample of matched pairs. We also obtain individual treatment effects by applying GRF with the matched sample.

We find adopting one-click buying is effective in lifting customer purchases and does so by making treated customers purchase more often and more items. The impact of joining one-click buying on customer purchases is economically

Table 1.11: TWFE Identification Check

	Purchase Amount ($)	Order Frequency	Items Purchased
	(1)	(2)	
Preperiod 1	1.008	0.069	-0.406
	(5.544)	(0.145)	(0.433)
Preperiod 2	5.038	0.195	0.320
	(7.597)	(0.154)	(0.435)
Preperiod 3	5.019	0.118	0.202
	(5.348)	(0.137)	(0.425)
Preperiod 4	4.877	0.084	0.487
	(4.075)	(0.141)	(0.587)
Preperiod 5	1.007	-0.164	-0.105
	(4.413)	(0.155)	(0.411)
Preperiod 6	4.558	-0.040	0.238
	(5.793)	(0.151)	(0.507)
Preperiod 7	-7.791	-0.408**	-0.991
	(5.576)	(0.150)	(0.561)
Preperiod 8	2.900	0.081	-0.006
	(3.457)	(0.141)	(0.328)
Preperiod 9	-0.272	0.050	-0.648
	(4.254)	(0.143)	(0.745)
Preperiod 10	2.043	-0.079	0.284
	(3.904)	(0.126)	(0.896)
Preperiod 11	-0.474	-0.081	-0.333
	(2.697)	(0.101)	(0.217)
Period fixed effects	Yes	Yes	Yes
Customer fixed effects	Yes	Yes	Yes
No. of customers	1,612	1,612	1,612
Observations	19,344	19,344	19,344

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Robust standard errors are clustered at the customer level and shown in parentheses.

significant, persistent over time, and heterogeneous across customers. Our findings are robust to potential confounding effects of self-selection and unobservables, different treated and control groups and different outcomes of purchase behavior.

To uncover the underlying mechanisms that lead to the behavioral changes, we leverage a variety of data including offline purchases, product returns, customer activity online, and purchases across product categories. Using these data, we explore channel switching, impulse buying, and customer engagement as possible explanations. Although multiple drivers can be at work, we suggest one-click ordering provides an avenue for the online retailer to improve customer engagement, which eventually leads to the positive changes in purchase behavior. We provide evidence consistent with this explanation based on micro-level data on customer activity online and purchase behavior across categories. We find no evidence for channel substitution and impulse buying.

Our findings offer three important managerial implications. First, this research presents evidence on the economic value of one-click buying for ecommerce companies. Because an increase in purchases as well as customer activity is expected for an ecommerce company after launching the technology, having management solutions in place to effectively and efficiently manage traffic to the website and fulfill orders so that customers have a positive experience is essential. Second, because companies are increasingly concerned about customer engagement, our findings illustrate the importance of investments and efforts toward improving customer experience through advanced features in ecommerce (e.g., seamless checkout process). Third, this study provides evidence on the heterogeneous treatment effects of one-click buying that can assist

managers in scoring and targeting customers and also in allocating resources across individual customers when designing marketing activity associated with one-click buying.

Because this research is the first attempt to quantify the effect of one-click buying on a variety of behavioral measures, a number of limitations should be acknowledged and perhaps addressed in future research. First, given that our context lacks random assignment into treatment and control groups, our identification strategy hinges on the unconfoundedness assumption. Our data allowed us to control for a rich set of observable measures when creating our matched pairs, thus increasing our confidence in the credibility of this assumption in our context. Unfortunately, we are unable to rule out with certainty any bias that might result from unobservable factors. Second, our study focused on a single retailer in a specific industry. Therefore, replication across other firms and industries as well as platforms would be needed to build empirical generalizations on this topic. With that in mind, we hope our approach provides a framework for further studies. Third, although we do not find any evidence for impulse buying, it's worth noting that the lack of evidence could be specific to our study context. Also, this could be because how we constructed measures of impulse buying based on customer behavior of product returns. Richer data is needed to conduct deeper analysis of possible explanations for the effects of one-click buying, which would be an interesting area for future research. Finally, we are unable to study how customer behavior at the focal retailer may change when competitors introduce their versions of one-click checkout. With competition in play, the effect of one-click buying on customer behavior remains unclear. Also it would be fruitful to explore the share of wallet for a possible explanation of one-click buying. We hope our work will generate further inter-

est in expanding our understanding of the impact of technology in a growing ecommerce landscape.

References

Angrist, Joshua D and Jörn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, NJ.

Ascarza, Eva (2018). "Retention futility: Targeting high-risk customers might be ineffective". In: *Journal of Marketing Research*, 55(1):80–98.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". In: *Annals of Statistics*, 47(2):1148–1178.

Athey, Susan and Stefan Wager (2019). "Estimating Treatment Effects with Causal Forests: An Application". In: *arXiv preprint arXiv:1902.07409*.

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

Baumgartner, Hans (2002). "Toward a personology of the consumer". In: *Journal of Consumer Research*, 29(2):286–292.

Bell, David R, Santiago Gallino, and Antonio Moreno (2020). "Customer supercharging in experience-centric channels". In: *Management Science*, 66(9):4096–4107.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). "How much should we trust differences-in-differences estimates?" In: *Quarterly Journal of Economics*, 119(1):249–275.

Bolton, Ruth N, Katherine N Lemon, and Peter C Verhoef (2004). "The Theoretical Underpinnings of Customer Asset Management: A Framework and Propositions for Future Research". In: *Journal of the Academy of Marketing Science*, 32(3):271–292.

Boone, Christophe and Walter Hendriks (2009). "Top management team diversity and firm performance: Moderators of functional-background and locus-of-control diversity". In: *Management Science*, 55(2):165–180.

Bower, Amanda B and James G Maxham III (2012). "Return shipping policies of online retailers: Normative assumptions and the long-term consequences of fee and free returns". In: *Journal of Marketing*, 76(5):110–124.

Breiman, Leo (2001). "Random forests". In: *Machine Learning*, 45(1):5–32.

Brynjolfsson, Erik, Yu Hu, and Duncan Simester (2011). "Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales". In: *Management Science*, 57(8):1373–1386.

Datta, Hannes, George Knox, and Bart J Bronnenberg (2017). "Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery". In: *Marketing Science*, 37(1):5–21.

De, Prabuddha, Yu Hu, and Mohammad S Rahman (2010). "Technology usage and online sales: An empirical study". In: *Management Science*, 56(11):1930–1945.

Digiday (2017). "End of an era: Amazon's 1-click buying patent finally expires." In: URL: https://digiday.com/marketing/end-era-amazons-one-click-buying-patent-finally-expires/.

Dutta, Ranjan, Sirkka Jarvenpaa, and Kerem Tomak (2003). "Impact of feedback and usability of online payment processes on consumer decision making". In: *ICIS 2003 Proceedings*, p. 2.

Fong, Nathan et al. (2019). "Targeted Promotions on an E-Book Platform: Crowding Out, Heterogeneity, and Opportunity Costs". In: *Journal of Marketing Research*, 56(2):310–323.

Forman, Chris, Anindya Ghose, and Avi Goldfarb (2009). "Competition between local and electronic markets: How the benefit of buying online depends on where you live". In: *Management Science*, 55(1):47–57.

Gallino, Santiago and Antonio Moreno (2018). "The value of fit information in online retail: Evidence from a randomized field experiment". In: *Manufacturing & Service Operations Management*, 20(4):767–787.

Häubl, Gerald and Valerie Trifts (2000). "Consumer decision making in online shopping environments: The effects of interactive decision aids". In: *Marketing Science*, 19(1):4–21.

Hui, Sam K et al. (2013). "The effect of in-store travel distance on unplanned spending: Applications to mobile promotion strategies". In: *Journal of Marketing*, 77(2):1–16.

Imbens, Guido W and Donald B Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, UK.

Iyengar, Raghuram, Young-Hoon Park, and Qi Yu (2022). "The Impact of Subscription Programs on Customer Purchases". In: *Journal of Marketing Research*. DOI: 10.1177/00222437221080163.

Keele, Luke (2015). "The statistics of causal inference: A view from political methodology". In: *Political Analysis*, pp. 313–335.

Kumar, Viswanathan and Denish Shah (2004). "Building and sustaining profitable customer loyalty for the 21st century". In: *Journal of Retailing*, 80(4):317–329.

Lemon, Katherine N and Peter C Verhoef (2016). "Understanding customer experience throughout the customer journey". In: *Journal of Marketing*, 80(6):69–96.

Magento (2017). "New Instant Purchase Checkout Boosts Sales." In: URL: https://magento.com/blog/magento-news/new-instant-purchase-checkout-boosts-sales.

Manchanda, Puneet, Grant Packard, and Adithya Pattabhiramaiah (2015). "Social dollars: The economic impact of customer participation in a firm-sponsored online customer community". In: *Marketing Science*, 34(3):367–387.

Narang, Unnati and Venkatesh Shankar (2019). "Mobile App Introduction and Online and Offline Purchases and Product Returns". In: *Marketing Science*, 38(5):756–772.

Palmer, Jonathan W (2002). "Web site usability, design, and performance metrics". In: *Information Systems Research*, 13(2):151–167.

Parboteeah, D Veena, Joseph S Valacich, and John D Wells (2009). "The influence of website characteristics on a consumer's urge to buy impulsively". In: *Information Systems Research*, 20(1):60–78.

Piron, Francis (1991). "Defining impulse purchasing". In: *Advances in Consumer Research*, 18:509–514.

Prins, Remco and Peter C. Verhoef (2007). "Marketing Communication Drivers of Adoption Timing of a New E-Service among Existing Customers". In: *Journal of Marketing*, 71(2):169–183.

Rafieian, Omid and Hema Yoganarasimhan (2021). "Targeting and privacy in mobile advertising". In: *Marketing Science*, 40(2):193–218.

Rejoiner (2017). "How Valuable is Amazon's 1-Click Patent? It's Worth Billions." In: URL: `http://rejoiner.com/resources/amazon-1clickpatent`.

Ridgway, Nancy M, Monika Kukar-Kinney, and Kent B Monroe (2008). "An expanded conceptualization and a new measure of compulsive buying". In: *Journal of Consumer Research*, 35(4):622–639.

Rossi, Peter E, Robert E McCulloch, and Greg M Allenby (1996). "The value of purchase history data in target marketing". In: *Marketing Science*, 15(4):321–340.

Shankar, Venkatesh (2018). "How Artificial Intelligence (AI) Is Reshaping Retailing". In: *Journal of Retailing*, 94(4):6–11.

Shannon, Claude Elwood (1948). "A mathematical theory of communication". In: *The Bell System Technical Journal*, 27(3):379–423.

Shehu, Edlira, Dominik Papies, and Scott A Neslin (2020). "Free shipping promotions and product returns". In: *Journal of Marketing Research*, 57(4):640–658.

Shi, Savannah Wei and Jie Zhang (2014). "Usage experience with decision aids and evolution of online purchase behavior". In: *Marketing Science*, 33(6):871–882.

Simester, Duncan, Artem Timoshenko, and Spyros I Zoumpoulis (2020). "Efficiently Evaluating Targeting Policies: Improving on Champion vs. Challenger Experiments". In: *Management Science*, 66(8):3412–3424.

Stern, Hawkins (1962). "The significance of impulse buying today". In: *Journal of Marketing*, 26(2):59–62.

Stuart, Elizabeth A (2010). "Matching methods for causal inference: A review and a look forward". In: *Statistical Science*, 25(1):1–21.

Thaler, Richard (1985). "Mental accounting and consumer choice". In: *Marketing Science*, 4(3):199–214.

Verhoef, Peter C et al. (2009). "Customer experience creation: Determinants, dynamics and management strategies". In: *Journal of Retailing*, 85(1):31–41.

Wager, Stefan and Susan Athey (2018). "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association*, 113(523):1228–1242.

Wagner, R Polk and Thomas Jeitschko (2017). "Why Amazon's '1-Click' Ordering Was a Game Changer". In: *Knowledge@Wharton*.

Wang, Kitty and Avi Goldfarb (2017). "Can offline stores drive online sales?" In: *Journal of Marketing Research*, 54(5):706–719.

Wang, Rebecca Jen-Hui, Edward C Malthouse, and Lakshman Krishnamurthi (2015). "On the go: How mobile shopping affects customer purchase behavior". In: *Journal of Retailing*, 91(2):217–234.

Witman, Allison et al. (2019). "Comparison group selection in the presence of rolling entry for health services research: rolling entry matching". In: *Health services research*, 54(2):492–501.

Xu, Kaiquan et al. (2016). "Battle of the channels: The impact of tablets on digital commerce". In: *Management Science*, 63(5):1469–1492.

CHAPTER 2

THE IMPACT OF ECOMMERCE-SAMPLING ON CONSUMER BEHAVIOR

2.1 Introduction

The share of ecommerce sales in total retail sales has been increasing in the US for the past few decades. In 2021, ecommerce sales accounted for about 15.3% of total retail sales and are expected to reach 23.6% by 2025 (Emarketer 2021). Because ecommerce will continue to grow in the foreseeable future, ecommerce companies constantly explore new ways of engaging their customers.

Despite its growing prominence in the retail landscape, ecommerce falls behind brick-and-mortar retail in at least one aspect, namely, communicating information about product quality and limitations of delivering that information have significant effects in the consumer decision process (Nelson 1970). Specifically, the lack of physical access to trying products before purchase in ecommerce increases uncertainty and thereby leads to friction in the decision process. This friction is particularly significant for physical experience products such as food, beverages, apparel, and so on which not only have search attributes that can easily be shared and conveyed to consumers online but also experiential attributes, which require physical inspection and cannot be communicated convincingly, such as taste, smell, feel, and fit (e.g., Lal and Sarvary 1999).

To overcome this shortcoming and help customers shop online, retailers have been implementing a diverse set of strategies that are aimed toward enabling the try-before-you-buy (TBYB) experience. Retailers have been enabling

the TBYB experience to customers in ecommerce in the comfort of their homes. As part of a business service, for example, most retailers now offer free shipping and return policies that allow customers to try products at home and return products that do not fit. Moreover, TBYB experience can take place both in store and online as part of a business service. In store it can materialize in the form of appointment shopping or showrooming, whereas online, it can transpire via virtual shopping or augmented reality that are tailored to communicate the experiential attributes of physical products ranging from apparel to furniture. Using Wayfair's augmented reality-enabled apps, for example, customers can gauge how a piece of furniture would fit in their living places. Similarly, using Sephora's Virtual Artist feature, customers can see how a beauty product would look on their face before placing an order from the online store.

A relatively new implementation of TBYB in ecommerce has emerged as well. Unlike the TBYB examples described above, this approach is a marketing promotion and can be described as the online version of the traditional product sampling which is also known as online or ecommerce product sampling (e.g., Forbes 2020). It has gained substantial attention in practice and is now considered one of the top three marketing tactics among marketers (e.g., Brandshare 2021). To implement TBYB as an online product sampling campaign, retailers and consumer brands send samples of selected products to online customers so that they can try them at home before purchasing them. For example, Sam's Club introduced online product sampling to its members in 2017 by sending them a bag of sample products (e.g., Grocer 2017). Similarly, customers can sign up to receive product samples on the Dove website, which offers a revolving variety of shampoos, lotions, soaps, and deodorants.

Online product sampling has gained traction recently among retailers and brands, in particular due to the restrictions that accompanied the COVID-19 pandemic. To remedy for the limitations placed on trying products in stores and to strengthen their relationship with online shoppers, big brand retailers such as Walmart started sending them various bags of samples ranging from a beauty box to a college bag (Forbes 2020). Engaging with customers via online product sampling has several benefits. First, it allows customers to try products in the comfort of their homes, which can earn attention for the products and expand shoppers' knowledge about the products. Second, it can communicate to customers that the firm is thinking about and cares for them, and can foster relationships as well as inspire loyalty with existing customers. Finally, it provides a safe option to experience products because customers do not have direct contact with another human. However, to date, no academic research has empirically examined whether exposure to online product sampling has any meaningful impact on customers' subsequent behavior.

The objective of this paper is to fill this gap in the literature and measure the long-term effect of the TBYB experience in the form of online product sampling on customer behavior and to explore possible explanations for the effect. Providing answers for this topic is important because they would generate insights that can guide brands and firms in their decisions about introducing online product sampling as well as designing targeting strategies for promoting it.

To achieve our research objective, we conduct a study in close collaboration with a retailer in Asia that implemented online product sampling on its online store in May 2020. Using quasi-experimental data over a period of 13 months

before and after the launch of ecommerce sampling, we apply a two-step identification strategy that combines propensity score matching with the difference-in-differences (DiD) (Angrist and Pischke 2008) procedure and estimate the average treatment effect (ATE) on customers who ordered online sampling at the firm's online store. Furthermore, we obtain individual-level treatment-effect estimates by applying generalized random forests (GRF) (Athey, Tibshirani, and Wager 2019) and investigate the heterogeneity of the effects in a data-driven and non-parametric way.

We find online product sampling is effective in lifting purchases of the products featured via sampling. The effect on customer purchases is not only economically significant but also persistent over time. On average, customers who had exposure to the products via online sampling increased their monthly purchase amount of products included in the sampling by 7% over a period of six months post treatment, compared with a group of control customers. Online sampling increased purchase amount by making treated customers purchase more items per order (an increase of 4%). The impact of online sampling is stronger immediately after customers had exposure to the products in the sampling, and decays over time. Our findings are robust to potential confounding effects of self-selection and unobservables.

We also find substantial variation in the treatment effect. Specifically, the magnitude of the increase in purchase amount ranges from 4% to 17% post treatment. Online sampling had the largest impact on high-value customers (based on past purchases) who had a strong relationship with the firm prior to the treatment. Furthermore, we find a significant spillover effect of the treatment on customer purchases of other products at the brands featured in the online

sampling. We also find that the impact is not limited to the online channel, but rather expands to both online and offline channels.

We explain our findings by drawing on the literature on consumer behavior and relationship marketing. Guided by previous work, we argue that online product sampling may work mainly through generating positive affect, which occurs in several ways in online product sampling. First, sample products help reduce uncertainty about products and thereby leads to positive affect in the form of higher liking, trust, and purchase intention of the products. Second, product sampling generates positive affect due to the exposure effect of the free sample products. Finally, product samples create positive affect in the form of gratitude toward the firm, which is accompanied by a strong desire to reciprocate the benefit received. We provide evidence that these psychological forces jointly translate into higher demand for the retailer.

The remainder of the paper is organized as follows. In §2.2, we discuss different types of the TBYB experience and the related literature. We describe our research setting and data in §2.3. In §2.4, we discuss our empirical methodology. We present our findings and discuss possible explanations for the effect in §2.5 and §2.6, respectively. We conclude in §2.7.

2.2 Related Literature

While more brands and firms are experimenting with the TBYB experience, implementations of TBYB for physical products typically vary on two dimensions: (1) how TBYB is being designed and operationalized and (2) which shopping channel is being used to enable TBYB. Firms have implemented the TBYB ex-

perince to customers as part of a business service or as a marketing promotion through the online or offline shopping channel. Hence, four types of the TBYB experience exists in practice, and are summarized in Table 2.1. In what follows, we describe each TBYB implementation in detail and discuss the related literature.

Table 2.1: Examples of TBYB for Physical Products

Operationalization	Channel	Example	Article
Business Service	Offline	Stores	Bell, Gallino, and Moreno (2018)
		Showrooms	Bell, Gallino, and Moreno (2020)
	Online	Free shipping & returns	Bower and Maxham III (2012)
			Shehu, Papies, and Neslin (2020)
		Augmented reality	Gallino and Moreno (2018)
Marketing Promotion	Offline	Product sampling	Smith and Swinyard (1983)
			Bawa and Shoemaker (2004)
	Online	Free shipping & returns	Bower and Maxham III (2012)
			Shehu, Papies, and Neslin (2020)
		Product sampling	Our study

The first type of TBYB is a business service that allows customers to test and experience products in physical stores. Although this approach is shared among many of brick-and-mortar retailers, some have gone a step further and introduced showrooms that are mainly dedicated to giving customers the physical space to engage with products before committing to a purchase. Examples include Bonobos, Clearly Contacts, Glossier, and Warby Parker. Bell, Gallino, and Moreno (2018) report that introducing showrooms for Warby Parker not only generates demand but also improves operational efficiency by increasing

conversion as well as decreasing returns. At the micro level, Bell, Gallino, and Moreno (2020) study how customer behavior changes after customers visit an experience-centric offline store and spend time engaging with products of a men's apparel retailer. They report that the visit experience causes customers to increase their spending as well as their shopping velocity but decreases the likelihood of returns.

Applications of TBYB experience are rich in ecommerce.[1] Nearly every retailer implements TBYB through free shipping and returns either as a service or a marketing promotion. Previous work in this domain has shown free shipping increases retailers' sales (e.g., Bower and Maxham III 2012); however, it also makes customers more likely to purchase high-risk products, which increases the overall return rate (e.g., Shehu, Papies, and Neslin 2020). With the proliferation of artificial intelligence, retailers are now also implementing TBYB as a service in the online channel by offering augmented reality-enabled applications. Sephora, for example, offers customers an augmented reality feature that enables them to try on makeup products virtually before purchasing. Online shoppers at Warby Parker can get access to all frames and try them on virtually to see instantly how they look on them. Relatedly, previous work has shown offering customers fit information via virtual tools at an online apparel retailer increases conversion rates as well as order value. Furthermore, such features make customers to expand their purchases to multiple categories as well as increase their likelihood of repeat purchases at the retailer (e.g., Gallino and Moreno 2018).

[1] Aside from physical products, which are the focus of our study, TBYB has also been implemented in the online context with digital or information products such as software and games. For these type of products, the focus has been on comparing the effects of different free sample promotions (e.g., Lee and Tan 2013) and designing the optimal levels of free samples (e.g., Cheng and Liu 2012; Li, Jain, and Kannan 2019) to increase firm profits.

As a marketing promotion in the offline channel, TBYB has a long history in marketing. Also described as product sampling, this promotion has been applied both as in-store sampling (e.g., Smith and Swinyard 1983; Lammers 1991) and as mail-delivery sampling (e.g., Scott 1976; Bawa and Shoemaker 2004) in the offline context, mostly with food products and beverages. Previous studies in these domains have shown product sampling has a positive effect on immediate (Lammers 1991) as well as long-term (Bawa and Shoemaker 2004) sales outcomes.

In this study, we focus on a relatively new implementation of TBYB for physical experience products in ecommerce as a marketing promotion. We study how exposure to product sampling of physical products in ecommerce influences customer behavior, which has not been reported in the literature. The closest work to ours is Lin, Zhang, and Tan (2019), who study the impact of TBYB as a marketing promotion for physical products on product ratings in ecommerce. They report that the TBYB marketing promotion increases the ratings of the products included in the promotion. Our study period includes the peak of the COVID-19 pandemic, which offers a unique opportunity to examine the effect of TBYB as a marketing promotion for physical products when customers could not try and experience products in physical stores.

We contribute to the above literature in three ways. First, we study how implementing TBYB for physical products in ecommerce as a marketing promotion affects customer behavior in the long term in contrast to past literature that focuses on the implementation of TBYB as part of a business service. Second, unlike past literature on free sample promotions, our study employs rich micro-level data, allowing us to measure the spillover effects of online product

sampling. We empirically show that online sampling affects customers' long-term behavior significantly not only for products featured in product sampling but also for other products that are not included in the promotion but belong to the brands featured in online sampling. We also document that the impact of online sampling is not limited to the online channel, but rather expands to both online and offline channels. Finally, unlike past literature on in-store or mail-delivery sampling where customers could try products in multiple ways, our research involves customers who had an option to test physical products only through online sampling before making a purchase. This allows us to measure the causal effect of online product sampling on customer purchases more precisely.

We also contribute to a growing literature on the effects of various marketing interventions on customer behavior, using data from field experiments or quasi-experiments (e.g., P. Manchanda, Packard, and Pattabhiramaiah 2015; Datta, Knox, and Bronnenberg 2017; Narang and Shankar 2019; Bell, Gallino, and Moreno 2020). A limited number of papers have examined heterogeneous treatment effects using machine-learning methods (e.g., Ascarza 2018; Fong et al. 2019; Simester, Timoshenko, and Zoumpoulis 2020; Rafieian and Yoganarasimhan 2021). Our paper adds to this stream of research by offering an application that combines machine-learning methods with well-established econometric methods to a marketing-related problem.

2.3 Research Setting and Data

To study the impact of online product sampling on customer behavior, we collaborated with a retailer in Asia that prefers to remain anonymous. The retailer specializes in personal care products and sells a wide range of consumer goods at both brick-and-mortar and online channels. The retailer launched a product-sampling campaign on its website in May 2020 for about one month. Because the COVID-19 pandemic was at its peak level during this time period, the retailer had limited operations in its offline stores. Furthermore, customers were no longer allowed to touch samples or test products in physical stores, due to concerns about spreading the virus. To mitigate some of the harmful consequences of the pandemic on customers, the retailer launched the online product-sampling campaign as a response to the limitations in trying products in offline stores.

The introduction of the online sampling was communicated to customers through mass emails and on the website, and no specific targeting was involved. It was offered to customers at no cost and was intended to increase engagement with customers during a challenging time. Importantly, the product sampling was unconditional, insofar as no purchasing was required for customers to order the free samples online. The campaign offered customers free samples of 10 select products in the personal-care category and came in travel-sized packages. The 10 products featured in the online sampling had already been on the market and were readily available in the retailer's offline and online stores. Because the primary purpose of online sampling was to engage with existing customers, the management team had decided on the sample products based on the seasonal-

ity and popularity of the products.[2]

Our data span a period of 13 months, starting from November 2019 to November 2020. They include a random sample of 8,730 customers who ordered the product sampling during the campaign period on a voluntary basis.[3] They constitute the treatment group in our analysis. For the purpose of comparison, we also obtained a random sample of 24,774 customers who visited the ecommerce website during the campaign period but decided not to place the order of sampling. In this way, we ensure the control group was aware of the online sampling but did not place the order, due to personal preferences rather than no awareness of the campaign. The retailer did not target treated customers with different promotions and communications throughout the data period.

The data consist of three parts: transaction data of customer purchase and return behavior, clickstream data of customer activity, and demographic data. The transaction data contain detailed information about each order made by a customer, including when a customer purchased a product and how much she paid for it. The data also include information on products returned and the brands of products purchased and returned. The online clickstream data contain detailed individual-level information on each visit to the website and customer activity upon visit, including when a customer visited the website and which pages (and how long) she viewed on the website. Our data also contain demographic characteristics of customers, including age and gender.

[2]Our analysis at the product level revealed that the products featured in online sampling were relatively new, expensive, and popular. Prior to the online-sampling campaign, the retailer offered about 7,000 items to customers and the sample products, on average, ranked at 36%, 90%, and 93% in terms of the elapsed time since launch, price, and sales, respectively.

[3]Because of the non-disclosure agreement we have with the collaborating firm, we are unable to disclose the total number of customers who responded to online product sampling during the campaign period.

We conduct the analysis at the customer-month level. Using transaction data, we define two outcome measures associated with customer purchases. Because we are primarily interested in assessing how effective online sampling is in lifting sales, our primary measure is the amount a customer spends per month (purchase amount).[4] We also consider the number of items per order (order size). Because online product sampling could have an impact on customer behavior not only online but also offline, both measures are based on customer purchases combined from both channels and are constructed at the customer-month level.

We also characterize the variety in purchase behavior with several metrics. First, we classify a product a customer purchased as focal versus other product on the basis of whether that product was featured in product sampling or not: (1) purchase of focal products and (2) purchase of other products. Second, we classify purchases of other products based on whether the product's brand was featured in the product sampling or not: (1) purchase of other products at the brands featured in the sampling (brand products) and (2) purchase of other products in other brands (other products). Similar to our primary outcome measures, we construct these measures with both purchase amount and order size. Using these measures allows us to decompose purchases into their components and explore how customers changed their purchase behavior upon product sampling.

[4] All transactions were recorded in the currency of the country in which the headquarters of the company was located. We converted purchase amount to US dollars using the average exchange rate over the data period.

2.4 Empirical Framework

In this section, we discuss our identification strategy and describe the treated and control groups. We then discuss the details of our econometric approach for estimating the ATEs on the treated and the GRF procedure, which we employ to obtain individual-level treatment effects.

2.4.1 Identification Strategy

Our main objective is to identify whether exposure to product sampling from an online retailer causes changes in customer behavior and examine the heterogeneity of the effect. The challenge we face is that exposure to product sampling is not random, but rather self-determined, which can threaten identification if outcomes and exposure to product sampling are simultaneously correlated with time-constant and time-varying unobservable differences between the treated and control groups. This correlation may exist to some degree in our context, resulting in potential selection bias. For instance, customers who made fewer purchases before the launch of product sampling might be more likely to respond, because they might expect to benefit more upon experiencing products through online sampling.

The ideal approach to identifying the causal effect would be a randomized experiment, which would involve sending free samples to customers selected at random from the firm's customer base and then comparing their subsequent behavior with a control group that did not receive the free samples. This strategy is practically challenging in our context, because sending a few samples to ran-

domly selected customers whose interest in trying those products is uncertain is simply cost prohibitive. Another reason a random experiment is not preferable in our context is its shortcoming in identifying long-term treatment effects post treatment. Because running an experiment for longer than a few months becomes increasingly difficult and costly, studying the long-term treatment effects becomes challenging.

An alternative approach, albeit less than ideal, would be randomizing among customers the eligibility of ordering the set of free samples, which would result in two groups of customers: the intention-to-treat and control groups. Although this strategy is feasible, it is not the one the partner retailer and we prefer for a few reasons. First, this strategy is not equivalent to a complete randomized control study, because the customers in the intention-to-treat group would also be self-selecting into treatment if they ordered the free samples. As such, identifying the causal effect of treatment would call for econometric adjustments that correct for bias emerging from self-selection. Second, randomizing eligibility can cause a backlash among customers and to false accusations of discrimination, which can devalue the firm's brand in the eyes of its customers.

Due to the practical challenges involved in randomization-based identification strategies, we resort to a quasi-experimental identification strategy that utilizes observational data. Our approach is to combine the selection-on-observables strategy with temporal data, which leads to the DiD identification strategy. This approach allows us to effectively minimize the threat to identification from potential selection bias in two steps. In the first step, we perform propensity score matching based on time-constant and time-varying observables that might influence both outcomes and the decision to respond to

product sampling. As a result of this procedure, we obtain pairs of treated and control customers who are similar to each other with respect to observable characteristics. In the second step, we take this sample of matched pairs and employ the DiD procedure.

Employing the DiD procedure allows us to account for time-constant unobservable differences between the treated and control groups without making any assumptions. To account for time-varying unobservables, however, we rely on the assumption that in the absence of treatment, the expected changes in the outcome measures of the treated would have been the same for the treated and control. Also known as the parallel-trends assumption in the literature on causal inference (Angrist and Pischke 2008), it constitutes the basis upon which the DiD method relies on and allows us to identify the causal effect of online sampling on subsequent customer behavior.

2.4.2 Treated and Control Groups

We define treated customers as those who responded to online product sampling during the one-month campaign period and had an option to experience physical products. For comparison, we define control customers as those who visited the ecommerce website during the campaign period but did not place the order for free samples and thus did not have an option to test physical products. This approach ensures the control group was aware of the campaign, and the reason they did not order the sample products could be attributed to their lack of interest in the campaign. For both groups, we define the six months before and after the campaign as the pre-treatment and post-treatment period,

respectively.

Before we report the effect of online sampling on customer purchases, we first evaluate the differences in customer purchases of products featured in product sampling (focal products) between customers in both groups. Panel A in Table 3.1 shows that, on average, treated customers spent $5.86 in the six-month post-treatment period, whereas those in the control group spent only $4.48 (diff. = 1.39, p-value < 0.001). Looking at the purchase amount in the pre-treatment period, we see that treated and control customers also had different purchase patterns prior to treatment. On average, treated customers spent $4.51 on products included in the product sampling, whereas non-treated customers spent $5.12 (diff. = -0.61, p-value < 0.05). Put together, purchase behavior differed considerably between the two groups, and treated customers spent substantially less than control customers before the treatment, but more after the treatment.

In addition, Table 3.1 shows the same pattern in the number of items purchased (order size). Together, these statistics suggest the control group is not comparable to the treated, and a naïve comparsion of the changes in customer purchases would suffer from selection bias.

2.4.3 Propensity Score Matching

To obtain a relevant control group that is similar to the treatment group with respect to time-independent and time-varying observables, we implement matching by estimating the propensity score, defined as a customer's propensity to order the product sampling online.

Table 2.2: Summary Statistics of Treated and Control Groups

	Treated	Control	Treated – Control	p-value
Panel A. Before Matching				
Pre-treatment period				
Purchase amount ($)	4.51	5.12	-0.61	0.03
Order size	0.12	0.16	-0.03	0.00
Post-treatment period				
Purchase amount ($)	5.86	4.48	1.39	0.00
Order size	0.19	0.15	0.04	0.00
Difference				
Purchase amount ($)	1.35	-0.64	1.99	
Order size	0.07	-0.01	0.08	
Observations	8,730	24,774		
Panel B. After Matching				
Pre-treatment period				
Purchase amount ($)	4.63	5.12	-0.51	0.18
Order size	0.13	0.14	-0.01	0.25
Post-treatment period				
Purchase amount ($)	5.85	4.43	1.41	0.00
Order size	0.19	0.13	0.06	0.00
Difference				
Purchase amount ($)	1.22	-0.69	1.91	
Order size	0.06	-0.01	0.07	
Observations	7,776	7,776		

Note: Observations are at the customer level, and summary statistics refer to the six-month total of each measure in the pre- and post-treatment periods.

We estimate the propensity scores using logistic regression with two sets of covariates. In the first set, we include measures that describe customers' overall relationship with the firm and their interest in its products and services. We include the average of customer purchases in the pre-treatment period: (1) elapsed time (days) since last purchase (recency), (2) number of purchases made (frequency), (3) number of items purchased among products featured in product sampling (order size: focal), (4) number of items purchased among products not included in product sampling (order size: others), (5) purchase amount ($) spent on products featured in product sampling (purchase amount: focal), (6) purchase amount ($) spent on products not included in product sampling (purchase amount: others), and (7) amount of discount ($) received (discount). We also include the average of online activity measures in the pre-treatment period: (8) number of website visits in the pre-treatment period (website visits) and (9) elapsed time (months) since having online account (tenure). In the second set of covariates, we capture the demographics of the customers by including their (10) age and (11) gender. Table 2.3 shows the summary statistics of the covariates and describes how the variables are operationalized.

We estimate the propensity score with a flexible function by creating interactions as well as second-order terms of the main 11 covariates. After estimating the propensity score model, denoted by $e(x;\beta)$ and parameterized by β, we follow Imbens and Rubin (2015) and transform it to obtain the linearized propensity scores, that is, the log-odds ratio:

$$l(x;\beta) = \ln\Big(\frac{e(x;\beta)}{1-e(x;\beta)}\Big).$$

We implement this transformation because it linearizes values on the unit interval and thus can improve the quality of the matching algorithm. We ask whether, for a treated customer i, a customer i' in the control group exists

Table 2.3: Covariates for Propensity Score Estimation

		Treated		Control	
Variable	Operationalization	Mean	Std. Dev.	Mean	Std. Dev.
Customer-firm relationship					
Recency	Elapsed days since last purchase	67.89	74.04	60.83	72.91
Frequency	Number of purchases made	4.59	5.69	6.55	8.02
Order size: Focal	Number of items purchased from sample	0.12	0.59	0.16	0.74
Order size: Others	Number of items purchased from others	17.79	31.32	23.26	37.64
Purchase amount: Focal	Purchase amount ($) spent on sample	4.51	21.38	5.12	24.94
Purchase amount: Others	Purchase amount ($) spent on others	203.50	386.81	225.22	446.12
Discount	Discount amount ($) received	51.52	139.55	86.22	239.19
Website visits	Average monthly visit sessions to the store	5.11	6.01	1.94	4.06
Firm tenure	Elapsed months since opening account	122.18	43.53	99.63	56.92
Socio-demographics					
Age		35.95	8.88	38.24	12.32
Gender	1 if female, 0 if male	0.98	0.14	0.86	0.34
Observations		8,730		24,774	

Note: Observations are at the customer level, and summary statistics refer to the pre-treatment means of each covariate in the treated and control groups.

such that the difference (in absolute value) in linearized propensity scores, $l(x_i;\beta) - l(x_{i'};\beta)$, is less than or equal to a threshold u. In our analysis, we focus on a threshold of $u = 0.05$ (Stuart 2010), meaning the difference in propensity scores is approximately less than 5%.[5] This matching algorithm results in 7,776 unique pairs that are closest to each other in the propensity scores.

[5]Results are robust with respect to the following values of u : $[0.05, 0.10, 0.15, 0.20, 0.20]$.

We evaluate the quality of the matching procedure in a few ways. First, we compare the distribution of the propensity scores between the treated and control groups and assess whether they are similar after matching. Figure 2.1 shows the density of the estimated propensity scores by treatment status, before and after matching. Before matching, the distributions share overlap but are significantly different from each other. After matching, the distributions closely resemble each other, and little bias seems to remain in the difference in the propensity scores between the groups.

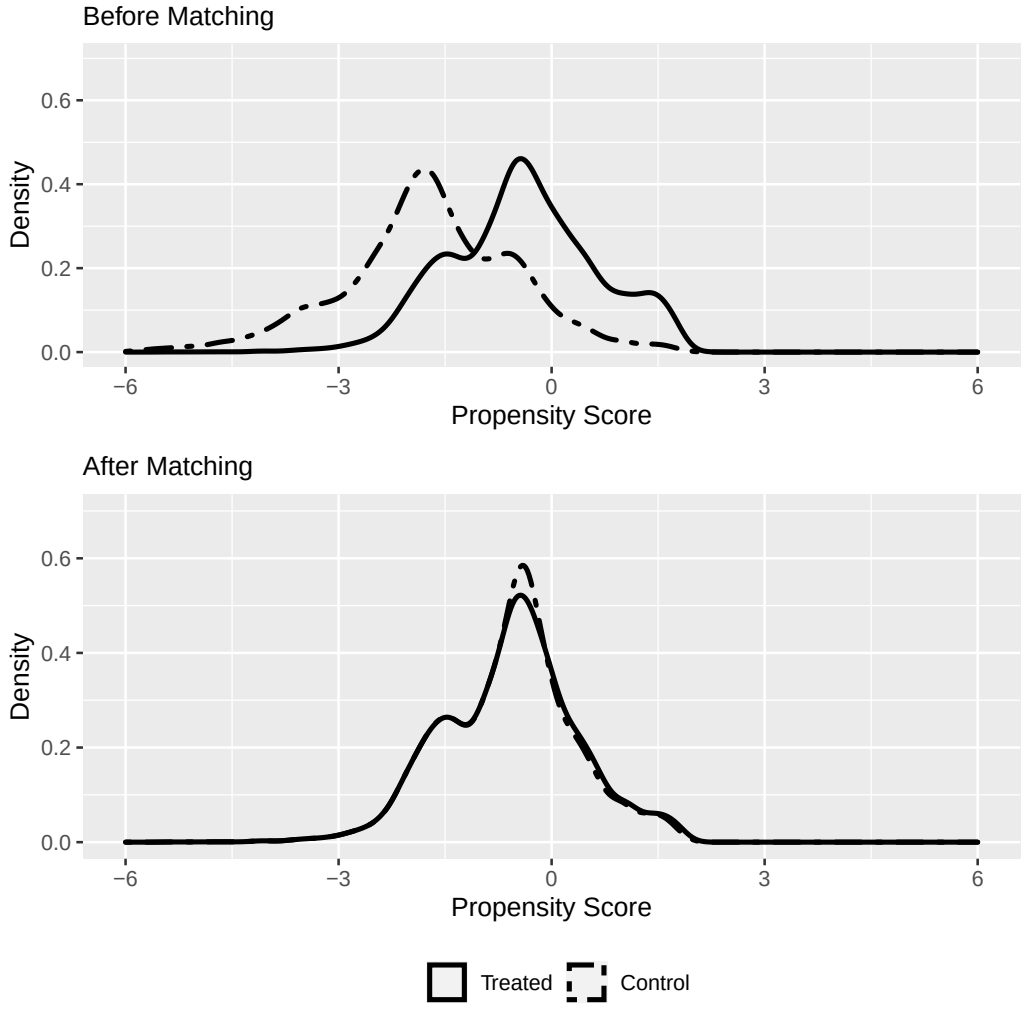

Figure 2.1: Distribution of the Propensity Score

We also assess whether matching achieves balance in the covariates across the treated and control groups. To that end, we follow (Austin 2009) and (Imbens and Rubin 2015) and examine the standardized differences in covariate means between the two groups. Figure 2.2 presents the standardized differences for each variable used in estimating the propensity scores. The figure shows that matching results in a substantial improvement in covariate balance. After matching, all of the normalized differences are below 0.1, which equals a degree of balance that one might expect in a completely randomized experiment (e.g., Stuart 2010; Imbens and Rubin 2015).

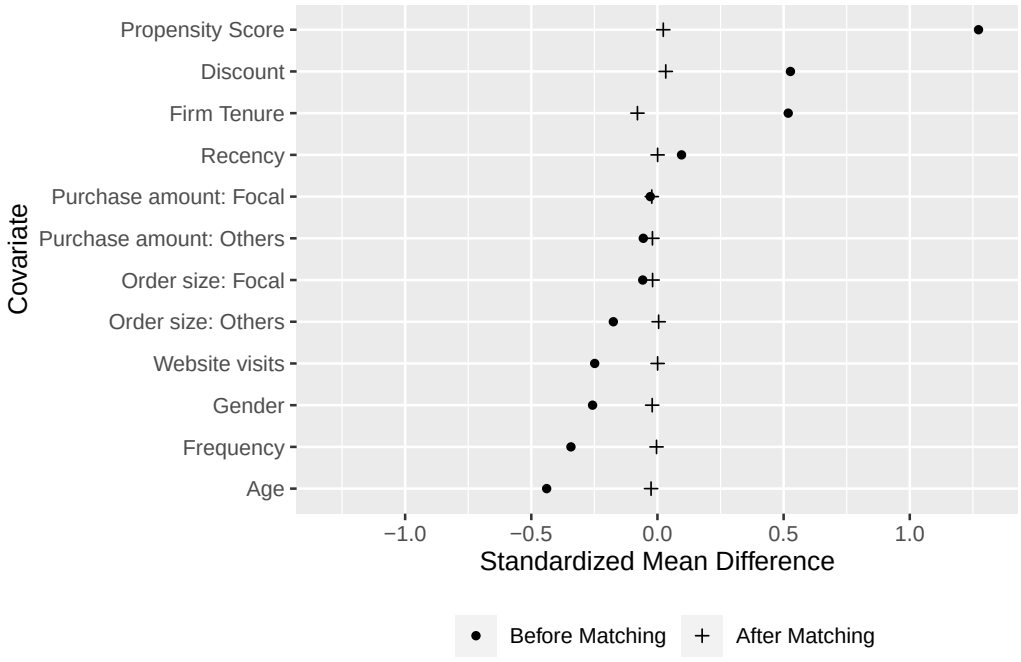

Figure 2.2: Covariate Balance

Panel B in Table 3.1 shows the summary statistics for the matched sample. The pre-treatment differences between the treated and control groups are smaller and no longer statistically significant. Taken together, using propensity score matching as the first step in our identification strategy, we successfully

address the differences in time-constant as well as time-varying observables between the treated and control groups.

2.4.4 Difference-in-Differences

The next step in our identification strategy consists of implementing the DiD method on the sample of 7,776 pairs obatined in the first step through matching. This step allows us to address the time-constant and time-varying unobservables that might simultaneously influence the decision to place the order of product sampling and our outcome measures.

We implement the DiD method with fixed effects for each individual, which effectively control for time-constant unobservable factors without making any assumptions. Controlling for time-varying unobservables, however, is less straightforward because doing so relies on the parallel-trends assumption. The challenge is that although this assumption is critical for properly executing the DiD method, its validity is fundamentally untestable because it involves a counterfactual.

Following the literature on causal inference, we assess the validity of the parallel-trends assumption by comparing the outcome trends between the treated and control groups in the pre-treatment period (Angrist and Pischke 2008). This approach is well established in the literature and rests on the idea that the assumption will fail if time-varying unobservables affect the treated and control customers differently and that the pre-treatment period can be used to gauge this possibility. Indeed, if the two groups had similar outcome trends before the treatment, such a finding is a valid indication that these trends would

have followed similar paths if the treatment had not occurred. The treatment, however, caused a shift in the counterfactual common trend for treated customers in the factual post-treatment period.

We evaluate the validity of the parallel-trends assumption first in a model-free manner by plotting the average outcome measures of the treated and control groups. We seek to achieve two goals in this exercise. First, to the extent that outcome trends are common prior to the treatment, we should see little indication of divergence in the averages of the outcomes in the months prior to the treatment, that is, $m \leq -1$. On the other hand, if ordering the free samples has any impact on the treated group, we should see a clear divergence in the averages of the outcomes in the post-treatment, starting at $m = 1$.

Figure 2.3 shows the plots over the data period for both outcome measures. The figure shows that our sample of matched pairs had similar purchase trends in the pre-treatment period, indicating the control group is a relevant comparison to the treatment group. Moreover, starting at $m = 1$, we see a sharp divergence from the common trend as the average outcomes of the treated group increased substantially compared with the period before the treatment. This finding suggests the treatment of ordering free samples online indeed caused changes in purchase behavior.

Next, we perform a model-based evaluation on the validity of the parallel-trends assumption by estimating the treatment effect of online sampling using the following DiD model:

$$\log(Y_{it}) = \theta_i + \sum_{m=-6}^{-1} \lambda_t \cdot 1(t = m) + \sum_{m=-6}^{-1} \beta_m \cdot W_i \times 1(t = m) + \epsilon_{it}, \tag{2.1}$$

where $\log(Y_{it})$ is the outcome measure by customer i in month t with natural log

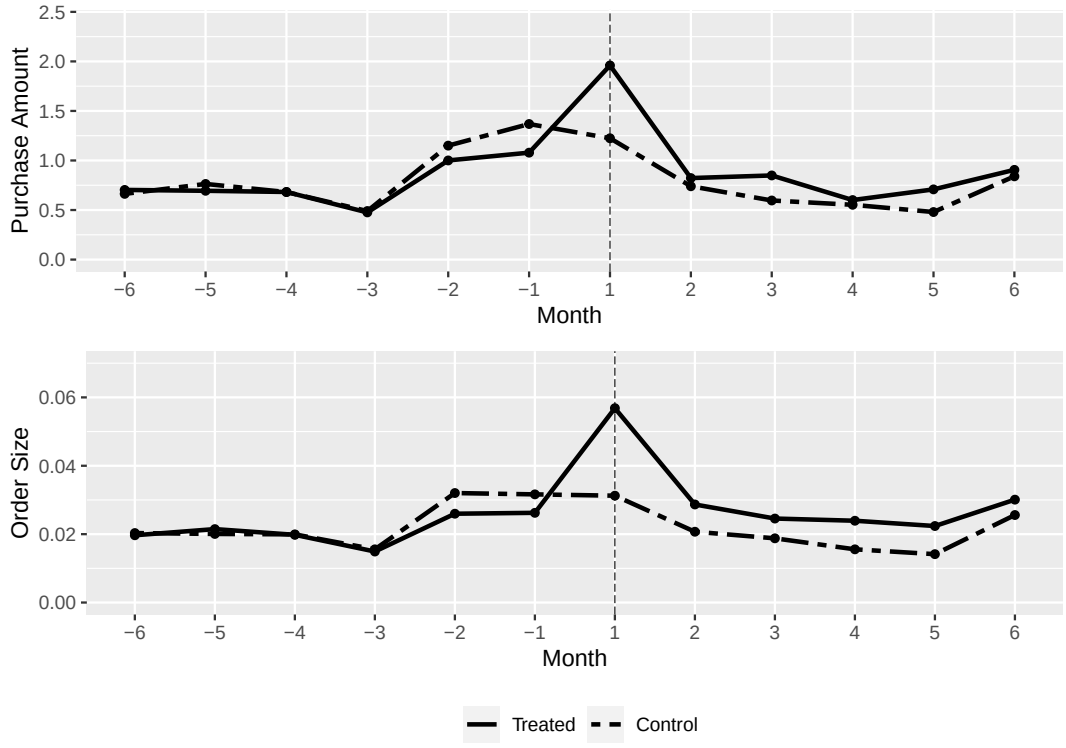

Figure 2.3: Comparison of Purchase Trends Between Treated and Control

transformation.[6] W_i indicates whether customer i belongs to the treatment or control group, and the indicator variables $1(t = m)$ are 1 if month t is m. θ_i is a customer fixed effect, λ_t is a month fixed effect, and ϵ_{it} is the error term. The two-way fixed-effects specification controls for time-invariant customer characteristics as well as common time trends and month-to-month fluctuations.

Our primary interest in this estimation is the parameter β_m. We normalize the first month ($m = -6$) as the baseline of 0 and estimate treatment effects over a period of six months before treatment. Parameter β_m captures the average difference in purchases between treated and non-treated customers relative to the baseline. To the extent that purchase trends are common prior to the treatment,

[6]To deal with 0 in the original scale, we added a constant of 0.01 before taking the natural log transformation. We perform robustness by using the outcomes in their original scale in §2.5.3.

all estimates of β_m in the months prior to the treatment ($m \leq -1$) should result in a null effect. We use robust standard errors clustered at the customer level to account for any serial correlation (e.g., Bertrand, Duflo, and Mullainathan 2004). Table 2.4 shows the results of estimating Equation (2.1). None of the estimates of the parameter β_m in the pre-treatment period are statistically significant, suggesting the matched pairs of treated and control customers had similar purchase trends before the sampling campaign.

In summary, our two-step identification strategy ameliorates concerns regarding self-selection bias and increases our confidence that we can identify the treatment effect of online sampling on subsequent purchases.

In what follows, we estimate the average effects of ordering free samples online on treated customers' purchase behavior over the entire six-month post-treatment period by employing the following DiD model:

$$\log(Y_{it}) = \theta_i + \sum_{m=-6}^{6} \lambda_t \cdot 1(t = m) + \beta \cdot W_i \times 1(1 \leq t \leq 6) + \epsilon_{it}, \tag{2.2}$$

where $\log(Y_{it})$ is the outcome measure by customer i in month t with natural log transformation. W_i indicates whether customer i belongs to the treatment or control group, and the indicator variables $1(\cdot)$ are 1 if the condition $(\cdot)$ is satisfied. θ_i is a customer fixed effect, λ_t is a month fixed effect, and ϵ_{it} is the error term.

After estimating Equation (2.2), we also investigate how the effects vary over time by obtaining estimates over three different periods: short term (within the first month of ordering, β_{1m}), medium term (between two and three months after ordering, β_{3m}), and longer term (between four and six months after ordering,

Table 2.4: DiD Identification Check

	Purchase Amount ($)	Order Size
	(1)	(2)
Treatment × Pre-month 5	-0.02	0.01
	(0.02)	(0.00)
Treatment × Pre-month 4	-0.01	0.00
	(0.02)	(0.00)
Treatment × Pre-month 3	-0.03	0.00
	(0.02)	(0.00)
Treatment × Pre-month 2	-0.03	-0.01
	(0.02)	(0.01)
Treatment × Pre-month 1	-0.05	-0.00
	(0.02)	(0.00)
Month fixed effects	Yes	Yes
Customer fixed effects	Yes	Yes
No. of customers	15,552	15,552
Observations	186,624	186,624

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Outcomes are in natural log terms. Robust standard errors are clustered at the customer level and shown in parentheses.

β_{6m}). To that end, we employ the following DiD model:

$$\begin{aligned}\log(Y_{it}) =&\theta_i + \sum_{m=-6}^{6} \lambda_t \cdot 1(t = m) + \beta_{1m} \cdot W_i \times 1(t = 1) \\ &+\beta_{3m} \cdot W_i \times 1(2 \leq t \leq 3) + \beta_{6m} \cdot W_i \times 1(4 \leq t \leq 6) + \epsilon_{it}. \end{aligned} \tag{2.3}$$

This specification allows us to examine whether the changes in purchase behavior are short lived or long lasting. In both specifications, we use customer and month fixed effects, which jointly control for time-invariant customer characteristics as well as common time trends and month-to-month fluctuations. We also implement robust standard errors clustered at the customer level to account for any serial correlation. We report the results of these analyses in §2.5.1 after explaining how we estimate the heterogeneity of the treatment effect.

2.4.5 Generalized Random Forests

We analyze the heterogeneity of the treatment effect by obtaining individual treatment-effect estimates through the GRF (Athey, Tibshirani, and Wager 2019) procedure. By leveraging machine-learning concepts, GRF offers a nonparametric statistical estimation method for causal inference in observational studies. The method is an extension of the causal forest method (Wager and Athey 2018), which is based on the classic random forest algorithm used for statistical learning (Breiman 2001).

The main idea of GRF is that the splitting criteria for growing individual trees are specifically designed to find partitions where treatment effects most differ. This feature allows us to find, in a data-driven way, the features that are most responsible for the heterogeneity in the treatment effects. Because the

method is based on random forests, it can accommodate nonlinear relationships among features and does not rely on functional-form assumptions for estimation. Without GRF, capturing nonlinearities in the data requires the creation of multiple interaction terms, and estimation becomes susceptible to the functional form.

Another useful feature of GRF is that as a byproduct of individual treatment effects, it yields doubly-robust ATEs. The ATE obtained from implementing GRF can then be used as a robustness check of other parametric estimation methods that have specific functional forms. In our application, we use the ATE from GRF as a robustness to our ATE based on the DiD procedure, which has a linear and additive form.

As a forest-based machine-learning application, GRF is vulnerable to overfitting, which is shared by all supervised learning methods. To minimize overfitting, we conduct hyper-parameter optimization using cross-validation and perform out-of-bag predictions. In particular, out-of-bag predictions identify for each example all the trees that did not use this example during training and makes predictions for it using only these trees. For more details about GRF, we refer readers to Athey, Tibshirani, and Wager (2019).

We leverage the panel structure of our data and apply the GRF procedure to the changes in the outcomes before and after the treatment, instead of using the outcomes themselves. This approach allows us to combine GRF with our DiD model and apply it to the matched sample. In this way, we do not compromise our identification strategy, which is based on the parallel-trends assumption, and employ GRF for estimation purposes only. In other words, instead of running regressions, we apply a non-parametric forest-based procedure to estimate

individual treatment effects, and the identifying assumptions are the same.

We define the outcome (Y_i) as the log-transformed change in purchase behavior for customer i as follows:

$$Y_i = \frac{1}{|T_A|} \sum_{t \in T_A} \log(Y_{it}) - \frac{1}{|T_B|} \sum_{t \in T_B} \log(Y_{it}), \tag{2.4}$$

where T_B and T_A denote the six months before and after treatment, respectively. To improve the performance of the causal forest, the developers of the algorithm recommend not including every available covariate and instead letting the forest search among covariates that are expected to cause heterogeneity in the effect. We follow this recommendation and include a subset of the covariates in estimating the propensity scores for the heterogeneity analysis. We include recency, frequency, focal monetary value, other monetary value, firm tenure, and website visits from the customer-firm relationship set. We also include age and gender in the analysis.

2.5 Findings

In this section, we report our findings for the ATEs on the treated customers (ATT) and discuss the heterogeneity of the effects. We also report the results of robustness tests with respect to functional-form assumptions.

2.5.1 Average Treatment Effects

Before we estimate Equation (2.3) and report the ATT by time period, we first estimate its simpler version by defining a single post-treatment period in Equation

(2.2). This approach gives us an overall idea about the ATT over the six-month post-treatment period. As Table 2.5 shows, we find that exposure to products through online sampling increased treated customers' purchase amount as well as order size. In the six months after treatment, treated customers on average increased their monthly purchase amount and order size by 7% and 4%, respectively.

Table 2.5: ATT Using DiD

	Purchase Amount ($)	Order Size
	(1)	(2)
Treatment × Post	0.07***	0.04***
	(0.01)	(0.01)
Month fixed effects	Yes	Yes
Customer fixed effects	Yes	Yes
No. of customers	15,552	15,552
Observations	186,624	186,624

Notes: $^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$. Outcomes are in natural log terms. Robust standard errors are clustered at the customer level and shown in parentheses.

We also investigate whether the effect varied by the channel (online vs. offline) through which customers made their purchases. In Table 2.6, we report the ATT based on the shopping channel. The results suggest the overall effect is almost equally divided and is statistically significant in both channels. The effect on online and offline purchase amount post treatment are 4% and 3%, respectively. Similarly, the effect on online and offline order sizes is 3% and 2%,

respectively.

Table 2.6: ATT by Channel Using DiD

	Online		Offline	
	Purchase Amount ($)	Order Size	Purchase Amount ($)	Order Size
	(1)	(2)	(3)	(4)
Treatment × Post	0.04***	0.03***	0.03***	0.02***
	(0.01)	(0.00)	(0.01)	(0.00)
Month fixed effects	Yes	Yes	Yes	Yes
Customer fixed effects	Yes	Yes	Yes	Yes
No. of customers	15,552	15,552	15,552	15,552
Observations	186,624	186,624	186,624	186,624

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Outcomes are in natural log terms. Robust standard errors are clustered at the customer level and shown in parentheses.

In Table 2.7, we report the ATT divided over time. The rows in the table are organized based on the DiD estimates from short (1 month) to long (4-6 months) terms. Our findings suggest the effect of exposure to product sampling from an online retailer is highest within the first month of the treatment. Importantly, we find the effect is persistent over time. Within the first month, the effect is 20% and 12% for purchase amount and order size, respectively, and drop to 4% and 2% between four and six months after treatment.

Table 2.7: ATT by Time Period Using DiD

	Purchase Amount ($)	Order Size
	(1)	(2)
Treatment × Short Post (1 month)	0.20***	0.12***
	(0.02)	(0.01)
Treatment × Medium Post (2-3 months)	0.05***	0.03***
	(0.01)	(0.01)
Treatment × Long Post (4-6 months)	0.04***	0.02***
	(0.01)	(0.01)
Month fixed effects	Yes	Yes
Customer fixed effects	Yes	Yes
No. of customers	15,552	15,552
Observations	186,624	186,624

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Outcomes are in natural log terms. Robust standard errors are clustered at the customer level and shown in parentheses.

2.5.2 Heterogeneous Treatment Effects

Using the estimates of the individual treatment effects obtained through the GRF procedure in §2.4.5, we now report the heterogeneity of the treatment effects. As discussed earlier, the averages of the individual treatment effects serve as a robustness check for the DiD estimates presented previously, which assume linear and additive treatment effects (Keele 2015). Table 2.8 reports the ATT obtained through GRF, which are similar to our findings in Table 2.5 from the DiD analysis.

Table 2.8: ATT Using GRF

	Purchase Amount ($)	Order Size
	(1)	(2)
Treatment × Post	0.07***	0.04***
	(0.01)	(0.01)
No. of customers	15,552	15,552
Observations	15,552	15,552

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Outcomes are in natural log terms. Bootstrapped standard errors are shown in parentheses.

We also find significant variation in the changes in purchase measures across customers as a result of ordering free samples online. Figure 2.4 shows how the individual-level treatment-effect estimates are distributed. The effects are positive for all of the treated customers, but the magnitude of the increase varies considerably, ranging from a mere 4% to 17% and 1% to 25% for purchase amount and order size, respectively. Table 2.9 reports the summary statistics of the individual treatment effects. Overall, these results illustrate the benefits of obtaining individual treatment-effect estimates by applying the GRF procedure.

Using the individual treatment effects, we next identify the subgroups of customers who have a stronger response to the treatment with respect to increase in purchase behavior. To do so, we relate the heterogeneous treatment effects to observed covariates and divide treated customers into two equal-sized groups based on their covariate values (i.e., below and above the median). We

Table 2.9: Heterogeneity of Treatment Effects Across Treated

	Mean	Std.Err.	Min	Max	N	$N_{\hat{\tau}>0}$	$N_{\hat{\tau}<0}$
Purchase Amount ($)	0.07	0.01	0.04	0.17	7,776	7,776	0
Order Size	0.04	0.01	0.01	0.25	7,776	7,776	0

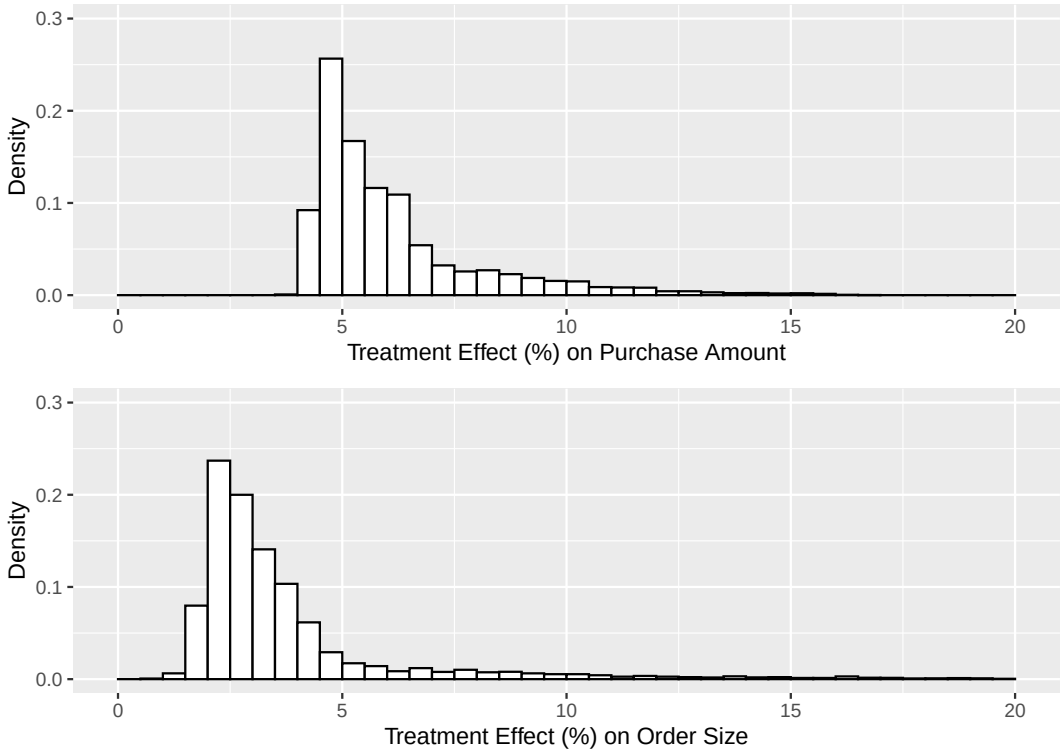

Figure 2.4: Distribution of the Treatment Effects

then compute the means of the personalized treatment-effect estimates across the two groups.

Figure 2.5 shows how the ATT differs among the groups with high and low covariate values. The figure demonstrates how purchase behavior changes as a function of the RFM measures: customers who purchased more recently, more frequently, and spent more in the pre-treatment period had a larger increase in purchases after having exposure to product sampling. Similarly, customers who had a longer tenure with the firm as well as those who benefited from more discounts had a larger treatment effect. In terms of the customer-firm relationship, this finding suggests the online product sampling had a stronger effect on those customers who already had a strong relationship with the firm and were essentially loyalists.

2.5.3 Robustness Checks

We analyze the robustness of our findings with respect to functional forms employed in our analysis. Specifically, we evaluate the sensitivity of our results to functional-form assumptions in the dependent variable and the response surface. We use the log transformation for our dependent variables in our main analysis for several reasons. First, purchase data are skewed, hence applying the log transformation provides a better fit. Second, by log-transforming dependent variables, we interpret the impact as percentage changes in outcomes. As such, comparing the impacts across different outcome measures becomes easier.

As part of our robustness checks with regards to functional forms, we es-

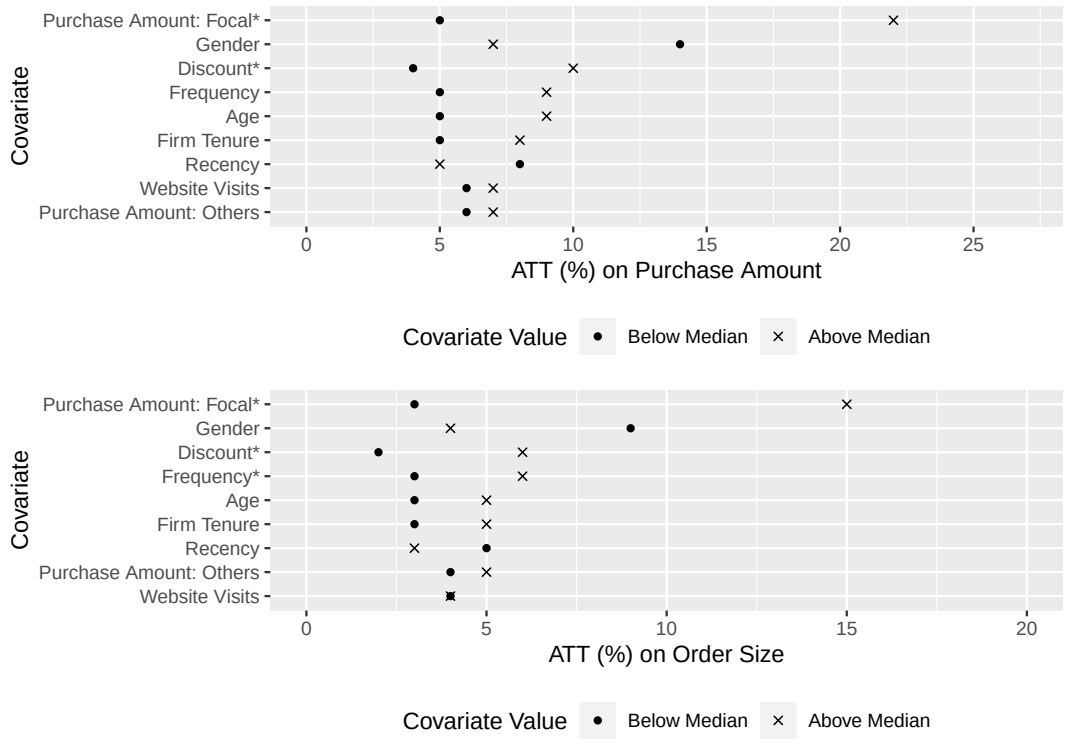

Figure 2.5: Conditional ATT (CATT) for High and Low Covariate Values

Note: * Indicates the CATT difference between high and low values of that covariate is statistically significant at 0.05.

timate our DiD model without log-transforming the outcomes. As shown in Tables 2.10 through 2.12, all estimates are positive and statistically significant, suggesting our main findings are robust in terms of functional forms of the outcomes. The results suggest an average monthly increase of \$0.32 in purchases of the sampled products among the treated customers after ordering the free samples in the long term. Another test with regards to the functional forms involves evaluating the linear and additive structure of the DiD estimation. As discussed in §2.4.5, the estimation via GRF allows us to relax this assumption, and as we discussed in §2.4.5, our findings are not influenced by this assumption.

Table 2.10: ATT Using DiD without Log Transformation

	Purchase Amount ($)	Order Size
	(1)	(2)
Treatment × Post	0.32***	0.01***
	(0.07)	(0.00)
Month fixed effects	Yes	Yes
Customer fixed effects	Yes	Yes
No. of customers	15,552	15,552
Observations	186,624	186,624

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Robust standard errors are clustered at the customer level and shown in parentheses.

2.6 The Role of Affect

Marketers use product sampling to introduce their products to new audiences, to strengthen relationships, and to increase engagement with existing customers as well as encourage sales of new products. One reason this product sampling can influence customer behavior is that it can create positive affect, a pleasant internal-feeling state that in turn can lead to a positive judgment and evaluation of the products sampled as well as the brands of those products. Among the different types of affect in consumer judgment and decision-making, integral affect is particularly relevant in product sampling (e.g., Cohen, Pham, and Andrade 2018).

Integral affect is the affective responses that are evoked through direct expe-

Table 2.11: ATT by Channel Using DiD without Log Transformation

	Online		Offline	
	Purchase Amount ($)	Order Size	Purchase Amount ($)	Order Size
	(1)	(2)	(3)	(4)
Treatment × Post	0.20***	0.01***	0.12*	0.00*
	(0.04)	(0.00)	(0.1)	(0.00)
Month fixed effects	Yes	Yes	Yes	Yes
Customer fixed effects	Yes	Yes	Yes	Yes
No. of customers	15,552	15,552	15,552	15,552
Observations	186,624	186,624	186,624	186,624

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Robust standard errors are clustered at the customer level and shown in parentheses.

rience with the object of judgment or decision. The pleasant feeling of tasting a wine or smelling a fragrance are examples of positive affective responses that give rise to integral affect. These affective responses are integral because they are elicited by features of the object that one experiences after direct contact. Studies have shown the existence of a strong relationship between integral affect and object evaluation to the extent that objects that generate positive integral affect are evaluated more favorably, and vice versa. Moreover, as a result of positive affect, product sampling positively influences consumers' belief and confidence in the product (e.g., Marks and Kamins 1988), perceived quality of the product, and evaluation of the brand and purchase intention (e.g., Kempf and Smith 1998). These positive attitudes toward the featured products can be expected to be larger immediately after exposure to the product sampling when

Table 2.12: ATT by Time Period Using DiD without Log Transformation

	Purchase Amount ($)	Order Size
	(1)	(2)
Treatment × Short Post (0-1 months)	0.82***	0.03***
	(0.19)	(0.01)
Treatment × Medium Post (2-3 months)	0.25**	0.01**
	(0.10)	(0.00)
Treatment × Long Post (4-6 months)	0.19*	0.01*
	(0.10)	(0.00)
Month fixed effects	Yes	Yes
Customer fixed effects	Yes	Yes
No. of customers	15,552	15,552
Observations	186,624	186,624

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Robust standard errors are clustered at the customer level and shown in parentheses.

consumers' direct experience with the products is more salient, compared with longer time periods after exposure (e.g., Cohen, Pham, and Andrade 2018).

Allowing customers to physically try and experience a sample of a product before purchasing it can create positive integral affect toward the products in several ways. First, mere exposure in product trials has been shown to generate positive affective responses toward the products (e.g., Kempf and Smith 1998). These positive affective responses have also been shown to be stronger for hedonic or experience products than for functional products (e.g., Kempf 1999). Furthermore, the fact that product trials are costless to the consumers can be a

reason to evoke positive affect. Indeed, Shampanier, Mazar, and Ariely (2007) show that free offers evoke positive affect and consumers use this affect as input for their decision-making process.

Product trial can also generate positive affect by reducing the uncertainty about product quality and fit. The reason is that compared with other sources of product information such as word of mouth or advertising, consumers are shown to be more responsive to direct information of products, in particular to information about their experiential attributes (e.g., Wright and Lynch 1995). They tend to overrate personal experience with a product and to discount indirect experiences (e.g., Marks and Kamins 1988; Kempf and Smith 1998). As product sampling reduces uncertainty, it helps eliminate an important source of friction in the consumer decision-making process and allows them to make more informed decisions (e.g., Biswas, Grewal, and Roggeveen 2010). Moreover, according to uncertainty-reduction theory, lower levels of uncertainty is linked to pleasant feelings and increased liking toward the object of interest (e.g., Berger and Calabrese 1974).

Aside from generating positive affect toward the products, product trial can generate positive affect toward the business and brands as well. When consumers receive a benefit from a business in the form of a small gift, courtesy, or even extra effort, they develop positive affect specifically by having feelings of gratitude. These feelings are accompanied by trust toward the business and brands and a strong urge to reciprocate the benefit received (e.g., Morales 2005; Dahl, Honea, and R. V. Manchanda 2005). Previous research has shown the urge to reciprocate leads to not only increased purchase intentions but also changes in consumer behavior toward the business and brands. For example, consumers

who received a free sample from an online retailer are shown to give higher ratings to those products due to reciprocity (e.g., Lee and Tan 2013). However, feelings of gratitude decay over time and gratitude-based reciprocal behavior tend to be stronger in the short term, in particular immediately after the benefit is received (e.g., Palmatier et al. 2009).

For consumers to engage in gratitude-based reciprocal behavior, the benefit received must be perceived as a benevolent act that shows the business is caring for the consumers and not simply trying to increase its own profit (e.g., Palmatier et al. 2009). Because online product sampling involves sending customers free samples so they can try them at home, it is likely to be perceived as a sign of goodwill of the business. Moreover, sending free samples during especially hard times such as a pandemic is likely to elevate this perception of goodwill and thereby lead to even higher levels of gratitude.

Guided by the aforementioned findings in the literature, we conjecture that the impact of online product sampling on customer behavior primarily works through generating positive affective responses among customers. To investigate whether this suggested explanation indeed plays a role in the treatment effects, we perform several analyses in which we utilize data on online activity, product returns, and customer purchases at the brand level.

First, we argue that if online product sampling generates positive affect toward the sample products by reducing uncertainty, one way this can manifest itself is through a reduction in online search activity related to the featured products. When customers are uncertain about product quality and fit, they will actively seek additional information to reduce such uncertainty, and experiencing the products first-hand likely gives customers enough information

so that the need for additional search might be reduced. We test this idea by utilizing our data on online activity related to sample products and estimate Equation (2.2) on the following measures: number of website visits, number of page views, and amount of time spent on the online store. As shown in Table 2.13, compared with control customers, treated customers' online activity related to sample products decreased by 3% on average post treatment across all three measures.

Table 2.13: ATT on Online Activity

	Website Visits	Page Views	Duration
	(1)	(2)	(3)
Treatment × Post	-0.03***	-0.03***	-0.03***
	(0.01)	(0.01)	(0.01)
Month fixed effects	Yes	Yes	Yes
Customer fixed effects	Yes	Yes	Yes
No. of customers	15,552	15,552	15,552
Observations	186,624	186,624	186,624

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Robust standard errors are clustered at the customer level and shown in parentheses.

An additional way that positive affect and uncertainty reduction toward sample products can manifest itself is through lower return rates. If customers are less uncertain about sample products due to trial, we expect to see a difference in return rates when we compare the returns of sample products with the returns of other products that were not featured in the sampling.[7] Before online

[7] As the retailer offered about 7,000 items to customers, we included the products in the

product sampling, the average monthly return rates for sample products and other products in the same categories were 6.6% and 6.1%, respectively. After online product sampling, return rates increased to 7.9% and 8%, respectively, meaning the return rate increased by 19% for sample products, but by 33% for other products. This finding suggests online product sampling was indeed effective in limiting the increase in return rates.

As discussed earlier, online product sampling can also generate positive affect toward the business and brands as feelings of gratitude, which would be accompanied by a strong urge to reciprocate. If this indeed plays a role, we should see online sampling lift not only sales of the sample products included in the sampling but also sales of other products from the brands featured in the sampling (brand products). In other words, we should see online product sampling generate spillover demand for the brands included in the sampling. In addition, we expect no changes in demand for other products in other brands (other products), because the effect of online product sampling has no reason to spill over to other brands. We investigate this idea by estimating Equation (2.2) with customer purchases at the brand level. As Table 2.14 shows, compared with control customers, treated customers increased their purchase amount and order size for products of brands included in the sampling by 8% and 10% on average, respectively. However, treated customers did not change their purchases of products in other brands.

The role of affect as a potential mechanism in explaining the treatment effects can also be seen both in the temporal dynamics and the heterogeneity of the effects. Specifically, affect-based behavior is argued to be stronger in the short term, immediately after exposure to the stimuli. This observation is

product categories of the sample products for other products in our analysis.

Table 2.14: ATT on Brand and Other Products

	Brand Products		Other Products	
	Purchase Amount ($)	Order Size	Purchase Amount ($)	Order Size
	(1)	(2)	(3)	(4)
Treatment × Post	0.08*	0.10***	-0.01	0.01
	(0.04)	(0.02)	(0.02)	(0.02)
Month fixed effects	Yes	Yes	Yes	Yes
Customer fixed effects	Yes	Yes	Yes	Yes
No. of customers	15,552	15,552	15,552	15,552
Observations	186,624	186,624	186,624	186,624

Notes: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Outcomes are in natural log terms. Robust standard errors are clustered at the customer level and shown in parentheses.

indeed what we have reported in §2.5.1; that is, the effect is strongest in the first month after treatment and decreased as time passed. Regarding our findings of the heterogeneity of the effects, we posit that they are also consistent with the affect-generating mechanism. The reason is that previous work has demonstrated launching new promotions or services has a diminished impact on customers who had more experiences with the firm prior to receiving the treatment (e.g., Bell, Gallino, and Moreno 2020). Therefore, customers who are already loyal have a higher threshold to be motivated to increase their spending with the firm. Therefore, if online product sampling had been perceived as yet another promotional offer, we would have seen a similarly diminished impact among loyal customers. However, our results suggest the opposite, and we argue the reason is the campaign's power in generating affective responses from

the customers.

Taken together, our analyses using granular data on online activity, product returns, and purchases at the brand level, as well as our findings regarding the temporal patterns and the heterogeneity of the treatment effect, jointly provide evidence for positive affect generation and uncertainty reduction as the mechanism of the effect of online product sampling on customer behavior.

2.7 Conclusions

The growing importance of ecommerce in the retail landscape has motivated retailers to employ new ways to enable the TBYB experience. A relatively new one that can serve this purpose is online product sampling, whereby a business or brand sends samples of physical products to consumers and allows them to try the products in the comfort of their homes. This practice has gained traction and is increasingly becoming popular among consumer brands and retailers, which naturally begs the question of its economic value for the business.

Through our partnership with an omni-channel retailer that launched online product sampling during the height of the pandemic, we utilize quasi-experimental data over a 13-month period and measure the causal effect of online sampling of physical products on customer behavior post treatment. Our two-step identification strategy leverages the longitudinal aspect of our data and is based on first creating matched pairs of treated and control customers and then employing DiD estimation. We also otain individual treatment effects by combining GRF, a nonparametric statistical procedure based on machine learning, with our DiD model for the matched sample.

We find online product sampling is effective in lifting customer purchases and does so by making treated customers purchase more items per order. The impact of ecommerce sampling on customer demand is economically significant, persistent over time, and is heterogeneous across customers. Furthermore, we find the impact spills over positively to brand demand and to both online and offline channels. Our findings are robust to potential confounding effects of self-selection and unobservables.

We explain these findings through the theoretical lens of an affect-generation mechanism. We argue the impact of online product sampling may work through generating positive affect in customers in multiple ways. Product sampling increases consumers' belief and confidence in the products as well as the evaluation of the brands featured, due to the exposure effect of the free sample products. Moreover, exposure to the free samples reduces uncertainty about product quality and fit, which not only allows customers to make more informed decisions but also increases liking for the products. Finally, receiving free samples gives rise to an affective response in the form of gratitude from customers. Feelings of gratitude are accompanied by a strong desire to reciprocate the benefits received, which ultimately leads to the positive changes in purchase behavior.

Our research presents evidence on the economic value of online sampling of physical products. Because this research is among the first attempts to quantify the long-term impact of online product sampling on customer demand, a number of limitations should be acknowledged and perhaps addressed in future research. First, given that our context lacks random assignment into treatment and control groups, our identification strategy hinges on the parallel-trends as-

sumption. We have shown evidence that this assumption is likely to hold in our context. However, we are unable to rule out with certainty any bias that might result from time-varying unobservable factors. Second, our study focused on a single retailer in a specific industry. Therefore, replication across other firms and industries would be needed to build empirical generalizations on this topic. Finally, our study context took place in the midst of a pandemic, which has its own merits in terms of examining the effect of online product sampling at a unique time. However, more research is needed to establish the generalizability of these findings to normal market conditions. With that caveat in mind, we hope our approach provides a framework for further studies. We hope our work will generate further interest in improving our understanding of the TBYB experience in a growing ecommerce landscape.

References

Angrist, Joshua D and Jörn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, NJ.

Ascarza, Eva (2018). "Retention futility: Targeting high-risk customers might be ineffective". In: *Journal of Marketing Research*, 55(1):80–98.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". In: *Annals of Statistics*, 47(2):1148–1178.

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

Bawa, Kapil and Robert Shoemaker (2004). "The effects of free sample promotions on incremental brand sales". In: *Marketing Science*, 23(3):345–363.

Bell, David R, Santiago Gallino, and Antonio Moreno (2018). "Offline showrooms in omnichannel retail: Demand and operational benefits". In: *Management Science*, 64(4):1629–1651.

— (2020). "Customer supercharging in experience-centric channels". In: *Management Science*, 66(9):4096–4107.

Berger, Charles R and Richard J Calabrese (1974). "Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication". In: *Human Communication Research*, 1(2):99–112.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). "How much should we trust differences-in-differences estimates?" In: *Quarterly Journal of Economics*, 119(1):249–275.

Biswas, Dipayan, Dhruv Grewal, and Anne Roggeveen (2010). "How the order of sampled experiential products affects choice". In: *Journal of Marketing Research*, 47(3):508–519.

Bower, Amanda B and James G Maxham III (2012). "Return shipping policies of online retailers: Normative assumptions and the long-term consequences of fee and free returns". In: *Journal of Marketing*, 76(5):110–124.

Brandshare (2021). "US The New Era of Sampling Begins in the Home." In: URL: https://www.brandshare.us/ecommerce-sampling.

Breiman, Leo (2001). "Random forests". In: *Machine Learning*, 45(1):5–32.

Cheng, Hsing Kenneth and Yipeng Liu (2012). "Optimal software free trial strategy: The impact of network externalities and consumer uncertainty". In: *Information Systems Research*, 23(2):488–504.

Cohen, Joel B, Michel Tuan Pham, and Eduardo B Andrade (2018). "The nature and role of affect in consumer behavior". In: *Handbook of Consumer Psychology*. Routledge, pp. 306–357.

Dahl, Darren W, Heather Honea, and Rajesh V Manchanda (2005). "Three Rs of interpersonal consumer guilt: Relationship, reciprocity, reparation". In: *Journal of Consumer Psychology*, 15(4):307–315.

Datta, Hannes, George Knox, and Bart J Bronnenberg (2017). "Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery". In: *Marketing Science*, 37(1):5–21.

Emarketer (2021). "US Ecommerce 2021." In: URL: https://www.emarketer.com/content/us-ecommerce-forecast-2021.

Fong, Nathan et al. (2019). "Targeted Promotions on an E-Book Platform: Crowding Out, Heterogeneity, and Opportunity Costs". In: *Journal of Marketing Research*, 56(2):310–323.

Forbes (2020). "US Forbes 2020." In: URL: https://www.forbes.com/sites/charlesrtaylor.

Gallino, Santiago and Antonio Moreno (2018). "The value of fit information in online retail: Evidence from a randomized field experiment". In: *Manufacturing & Service Operations Management*, 20(4):767–787.

Imbens, Guido W and Donald B Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, UK.

Keele, Luke (2015). "The statistics of causal inference: A view from political methodology". In: *Political Analysis*, pp. 313–335.

Kempf, Deanna S (1999). "Attitude formation from product trial: Distinct roles of cognition and affect for hedonic and functional products". In: *Psychology & Marketing*, 16(1):35–50.

Kempf, Deanna S and Robert E Smith (1998). "Consumer processing of product trial and the influence of prior advertising: A structural modeling approach". In: *Journal of Marketing Research*, 35(3):325–338.

Lal, Rajiv and Miklos Sarvary (1999). "When and how is the Internet likely to decrease price competition?" In: *Marketing Science*, 18(4):485–503.

Lammers, H Bruce (1991). "The effect of free samples on immediate consumer purchase". In: *Journal of Consumer Marketing*.

Lee, Young-Jin and Yong Tan (2013). "Effects of different types of free trials and ratings in sampling of consumer software: An empirical study". In: *Journal of Management Information Systems*, 30(3):213–246.

Li, Hongshuang, Sanjay Jain, and PK Kannan (2019). "Optimal design of free samples for digital products and services". In: *Journal of Marketing Research*, 56(3):419–438.

Lin, Zhijie, Ying Zhang, and Yong Tan (2019). "An empirical study of free product sampling and rating bias". In: *Information Systems Research*, 30(1):260–275.

Manchanda, Puneet, Grant Packard, and Adithya Pattabhiramaiah (2015). "Social dollars: The economic impact of customer participation in a firm-sponsored online customer community". In: *Marketing Science*, 34(3):367–387.

Marks, Lawrence J and Michael A Kamins (1988). "The use of product sampling and advertising: Effects of sequence of exposure and degree of advertising claim exaggeration on consumers' belief strength, belief confidence, and attitudes". In: *Journal of Marketing Research*, 25(3):266–281.

Morales, Andrea C (2005). "Giving firms an "E" for effort: Consumer responses to high-effort firms". In: *Journal of Consumer Research*, 31(4):806–812.

Narang, Unnati and Venkatesh Shankar (2019). "Mobile App Introduction and Online and Offline Purchases and Product Returns". In: *Marketing Science*, 38(5):756–772.

Nelson, Phillip (1970). "Information and consumer behavior". In: *Journal of Political Economy*, 78(2):311–329.

Palmatier, Robert W et al. (2009). "The role of customer gratitude in relationship marketing". In: *Journal of Marketing*, 73(5):1–18.

Grocer, Progressive (2017). "US Progressive Grocer 2017." In: URL: https://progressivegrocer.com/sams-club-ventures-e-sampling.

Rafieian, Omid and Hema Yoganarasimhan (2021). "Targeting and privacy in mobile advertising". In: *Marketing Science*, 40(2):193–218.

Scott, Carol A (1976). "The effects of trial and incentives on repeat purchase behavior". In: *Journal of Marketing Research*, 13(3):263–269.

Shampanier, Kristina, Nina Mazar, and Dan Ariely (2007). "Zero as a special price: The true value of free products". In: *Marketing Science*, 26(6):742–757.

Shehu, Edlira, Dominik Papies, and Scott A Neslin (2020). "Free shipping promotions and product returns". In: *Journal of Marketing Research*, 57(4):640–658.

Simester, Duncan, Artem Timoshenko, and Spyros I Zoumpoulis (2020). "Efficiently Evaluating Targeting Policies: Improving on Champion vs. Challenger Experiments". In: *Management Science*, 66(8):3412–3424.

Smith, Robert E and William R Swinyard (1983). "Attitude-behavior consistency: The impact of product trial versus advertising". In: *Journal of Marketing Research*, 20(3):257–267.

Stuart, Elizabeth A (2010). "Matching methods for causal inference: A review and a look forward". In: *Statistical Science*, 25(1):1–21.

Wager, Stefan and Susan Athey (2018). "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association*, 113(523):1228–1242.

Wright, Alice A and John G Lynch Jr. (1995). "Communication effects of advertising versus direct experience when both search and experience attributes are present". In: *Journal of Consumer Research*, 21(4):708–718.

CHAPTER 3

THE IMPACT OF AN ONLINE PLATFORM'S ALGORITHM CHANGE ON VIEWER BEHAVIOR

3.1 Introduction

The digital economy is essential in both consumers' and firms' daily activities. At the heart of this ecosystem are online platforms. By bringing together consumers and intermediaries online platforms allow exchanges that would otherwise not happen, thus they are key in delivering benefits to consumers and businesses. Online platforms serve distinct group of users and broadly can be distinguished based on the key activities that consumers perform on them. They offer varied services such as Internet search engines (e.g. Google, Bing, Yahoo), online market places (e.g. eBay, Amazon), video-sharing platforms (e.g. Vimeo, YouTube), music and video platforms (e.g. Spotify, Netflix), social networks (e.g. Facebook, Twitter), collaborative economy platforms (e.g. AirBnB, Uber), online gaming (e.g. Steam), review platforms (e.g. Tripadvisor, Yelp) etc.

While online platforms are hugely diverse in terms of the services they offer, a shared attribute is their increasing utilization of algorithms. Many platforms routinely deploy algorithms as decision makers, ranging from purposes as simple as choosing the layout of a page or the shape of a button, to decisions as important as which website should be shown first or which news article should be given priority.

These algorithms play a major role in today's Internet economy. They increasingly serve as gatekeepers because they determine visibility, sharing and

flow of information (Tufekci 2015). To name a few examples, Google's search ranking algorithm determines how websites should be ranked in response to a keyword search, thus is critical in generating traffic to websites. Amazon's Buy Box algorithm decides, for a given product being sold by many sellers, which of the sellers will be featured in the Buy Box on the product's landing page (Chen, Mislove, and Wilson 2016). Facebook's News Feed algorithmic decision maker decides whether a post shared by one of its users or a news article by a publisher is shown to other users or not.

A key characteristic of these algorithms is that details about their workings is usually not visible to the public. In addition, in many social or entertainment platforms such as Facebook and YouTube they are tailored to each individual, making them highly capable of customizing the content viewers receive. In some cases online platforms inform the public when their algorithmic decision maker undergoes through an important update. Google, for example, announced on January 21, 2011 that the growth of content farms and low-quality sites threatened users' trust in Google's search result and that it was going to develop some changes to its search results ranking algorithm.[1] Similarly, starting on August 6, 2013, Facebook has been announcing new features of its News Feed algorithm, which is the primary system through which users are exposed to content posted on the network, on its blog.[2]

This non-transparent gatekeeping capacity of algorithms poses a fundamental challenge for society in large: every tweak to these algorithms can have huge impacts on users, the platforms' intermediaries and other stakeholders. Google's update dubbed as Google Panda, for example, was released in Febru-

[1] https://googleblog.blogspot.com/2011/01/google-search-and-search-engine-spam.html
[2] https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed

ary 2011 and its effect went global in April 2011. The impact was particularly heavy on media companies that had a search advertising-heavy business model because traffic to their websites decreased dramatically after the change. Demand Media, an online content maker, lost about 34% of its shares within a couple of weeks after Google's Panda update went into effect (Swisher 2011).

Recent research in marketing and social sciences have reported that algorithmic manipulations can have a substantial impact on consumers' feelings and attitudes (e.g., Hauser et al. 2009; Kramer, Guillory, and Hancock 2014; Bakshy, Messing, and Adamic 2015), however much of their effects on users and intermediaries of platforms remains a mystery. In particular, the impacts of algorithms on users and other stakeholders can reach beyond the confines of the platform or website in which they are employed, and research into these potential effects is lacking. In this research we attempt to fill this gap by seeking answers to the following questions. What impact does an online platform's change of its algorithm have on its users' behavior towards an intermediary operating on that platform? Does the impact vary across specific dimensions of the content that the intermediary produces? Finally, what is the underlying mechanism of a potential impact?

Addressing these questions is challenging for a few reasons. To begin, one needs to keep abreast of the changes that a platform makes to its algorithm by constantly monitoring the platform and its announcements. However, even with information regarding the nature of the change to the platform's algorithm, analyzing its impact on users' behavior and the intermediaries' decisions outside of the platform is challenging because of data limitations. Only if one has the privilege of accessing proprietary data of a business that operates within the

platform, one can leverage the platform's disclosing of algorithm updates and compare the changes in measures of interest of that business and its customers with respect to the changes in the algorithm. Another challenge could be related to the type of data one has access to. Using observational data might not allow to isolate the causal effect due to selection bias and unobersvable confounders, whereas data from experiments that randomly assign users to treatment arms would be more conducive for studying questions such as ours.

We address these questions by focusing on the Internet's largest social networking platform: Facebook; and an online publisher that primarily produces content to be distributed on this platform but also operates its own website independent of the platform: Upworthy.com. Upworthy is an online publishing company that started in March 2012, and by the end of 2013, it was being called the fastest-growing media company in the world (Kamenetz 2013). The company was at the forefront of U.S. media from 2013 to 2015. At the core of its success was its idea of using attention-grabbing headlines, which since has been referred to as "clickbait".

Soon after Facebook created its blog for sharing News Feed updates, on December 2, 2013, Facebook announced first time that it was changing the way it ranked content in its News Feed algorithm in order to promote high-quality content (Kacholia and Ji 2013). While it was highly speculative what Facebook defined as high-quality content, Facebook News Feed manager Lars Backstrom acknowledged in an interview that the screening would be primarily at the source level rather than content itself. Practically, this meant that contents from certain publishers would be flagged as high-quality and promoted in the News Feed, while others would be demoted (Kafka 2013). The manager also stated

that no specific publisher would be targeted and did not disclose which publishers' content would be demoted.

The objective of this paper is to uncover how Facebook's decision to change its News Feed algorithm impacted users' behavior towards content Upworthy published on its website. Providing answers to this question is important in this era of Internet economy because it will allow us to better understand the implications algorithmic decision makers can have even beyond the control of the platforms that employ them and let managers prepare accordingly.

To achieve this objective we leverage data from a series of experiments that Upworhty conducted on its website over a period of 121 weeks. The purpose of the experiments was to test the effectiveness of headlines for a given story before publishing it, and find the one that would generate the highest clickthrough rate. Using data from experiments ensures the elimination of selection-bias and because the change to Facebook's algorithm occurred at a specific date and independent of the experiments, we are able to generate insights about the event's impact by comparing various measures of customer behavior as well as Upworthy's content management practices before and after that date.

A critical component of our analysis involves the classification of headlines. This is important for two reasons. First, we want to understand whether the impact varies with respect to the headline being clickbait vs non-clickbait. Second, Upworhty's clickbait type headlines might generate larger clickthrough rates compared to non-clickbait ones. Therefore we classify headlines in an automated manner using deep-learning methods for classification, and incorporate them in our empirical model. Our empirical model exploits the nested structure of the data and employs pre-post comparison. Experiments were conducted

weekly across 121 weeks, prior and after the change to the algorithm, and each one tested the effectiveness of multiple headlines.

We find Facebook's first change to its News Feed algorithm caused a significant decrease in clickthrough rates for headlines Upworthy tested on its website. Specifically, clickthrough rates decreased on average by about 0.8 percentage points in the period after the change took place. Considering that the average clickthrough rate was 2.1% before the change, Facebook's change to its News Feed algorithm caused almost a 40% reduction in the average clickthrough rate.

Our analysis also reveals that clickbait headlines generated higher clickthrough rates overall, but clickthrough rates declined similarly for clickbaits and non-clickbait headlines after the announcement. On the publisher's side, we find that it increased on average not only the number of experiments it conducted on a weekly basis but also the number of headlines it tested in each experiment.

A key way in which Facebook's algorithm change can impact users is by reducing the visibility of the publisher's content shared by other users. As visibility is key in receiving clicks (Bakshy, Messing, and Adamic 2015), reducing it will drop the clicks posts receive, which ultimately will lead to less traffic to the publisher's website. Because users who come to the publisher's website from Facebook are self-selected users who are more likely to be interested in the content, they are also more likely to spend more time on the website and click through other content. However, the change of the algorithm disrupts this self-selection process, hence results in decreased clickthrough rates. To test this mechanism, we leverage Google Trend data. Using trend data of keywords "Upworthy" and "Buzzfeed" as proxies for traffic to Upworthy.com and Buz-

zfeed.com we find evidence that Facebook's change to its algorithm indeed decreased interest to Upworthy.com.

Our paper is related to a few streams of research. We contribute to the growing literature in marketing and the social sciences that reports the impact of algorithmic decision makers on consumer's behavior as well as feelings. In a variety of settings previous research has shown that algorithmic manipulations ranging from product recommendations to website design can have a significant impact on consumers' behavior and feelings (e.g., Hauser et al. 2009; Chung, Rust, and Wedel 2009; Kramer, Guillory, and Hancock 2014). Our research adds to this literature by showing the impact of algorithms can reach beyond the platform or website in which they are enployed.

A rich literature on keyword search advertising in marketing have documented the effects of ad positions on consumer clicks and conversions (e.g., Agarwal, Hosanagar, and Smith 2011; Rutz and Trusov 2011; Chan and Y.-H. Park 2015). In a similar vein, the literature in ecommerce has shown that rank orders determined by algorithms in search engines has a significant impact on the clickthrough rates and conversion rates (e.g., Jerath et al. 2011; Ghose, Ipeirotis, and Li 2014). A common attribute of these studies is that they measure the impact of certain features of the algorithms and attempt to understand how to improve the effectiveness of these features. We add to this litearture by reporting the impact of the change to to the algorithm itself.

Additionally, we contribute to the nascent and growing literature on the causal inference of various marketing interventions on customer behavior using experimental data. Only a few papers in this area have used machine-learning methods (e.g., Ascarza 2018; Fong et al. 2019). Our paper adds to this stream

of research by combining traditional econometric methods with state-of-the-art machine-learning tools for causal inference.

The remainder of the paper is organized as follows. §3.2 describes our research setting and data. §3.3 discusses our empirical methodology. We present our findings in §3.4 and discuss the possible explanation for our findings in §3.5, and conclude in §3.6.

3.2 Research Setting and Data

In this section, we describe our research setting and data. In §3.2.1, we discuss the relationships online publishers may have with Facebook, and provide a description about Upworthy.com and the publishing strategy it employed during our data period in §3.2.2. In §3.2.3, we describe the data we obtained from the Upworthy research archive to study how the change in Facebook's News Feed algorithm impacted user behavior as well as publisher behavior.

3.2.1 Facebook and Publishers

Facebook as a platform is key for many online publishers because it generates not only ad revenue but also referral traffic to publishers' website. Between December 2011 and December 2014 Facebook became the number one source of traffic for many digital publishers, its share of total visits sent to publishers rose by 277.26%. In December 2014 Facebook sent 24.63% of the total visits publishers received, up from 6.53% in December 2011. Over a comparable time frame, Facebook's user base grew 60%, suggesting that the near 4x explosion in

traffic share was due to a far more engaged user base (Wong 2015).

The primary system that determines traffic to publishers' website from Facebook is the News Feed algorithm. Launched in September 2006, the algorithm is the key decision-maker through which users are exposed to content posted on the network. When users click on publishers' content - posted either by them or other users - they are diverted to the publishers' website, hence ranking higher on the News Feed at Facebook is critical for publishers [3]. Previous research has shown that the higher the link, substantially more likely it will be clicked on (Bakshy, Messing, and Adamic 2015). As such, an online publishing company whose business model heavily relies on the visibility it gains on social platforms, in particular Facebook, lives and dies by placement that is determined by the News Feed algorithm. Needless to say, the more traffic Facebook drives to publishers, the more reliant publishers become on the platform, and any changes to Facebook's News Feed algorithm has the potential to impact both users and publishers.

Upworthy is an online publishing company that started in March 2012 with a purpose of reaching large audiences with content on social and cultural issues that is entertaining to consume as well as compelling to share on social media. Since its inception the company has been meticulous about creating shareable content that appeals, so that they generate attention on online platforms, in particular Facebook. Towards this, the company has been publishing on the Internet content with topics like global poverty, domestic violence, drunken driving, gender bias, income inequality, AIDS, bullying, bigotry, and pediatric cancer.

[3]This was the case during our observation period. Facebook made other changes to News Feed since mid 2015. Facebook is now partnering with several publishers and allows users access content without directing them to publishers' websites.

Although the company's goal to reach large audiences with content that matters had noble intentions, its early strategy to achieve that goal were not different than contemporary digital intermediaries in the world of online publishing. Content aggregation rather than producing original content was the company's blueprint for publishing. Upworthy's writers (or curators as the company called them) searched the internet for stories with emotional salience and then republished them with compelling new headlines and tools to share the posts on social networks, mainly on Facebook.

Two main facts shed light on Upworthy's success in growing prominence within the Facebook universe. First, according to the company, within two years of its inception 78% of Americans on Facebook liked or had a friend who liked Upworhty's page. Second, people shared Upworthy posts at a rate that was nearly eight times the rate of the next comparable site. This suggests that Upworhy's reach primarily came from its core audience sharing links to everybody they knew on Facebook (Abebe 2014).

3.2.2 Publishing at Upworthy

The way in which Upworthy differed from other companies in online publishing, at least in its early years, was their approach in running experiments to find the optimal headline and image pair, defined as the package, to share a given story with. After finding interesting stories from around the Web, writers would repackage them with enticing headlines, and the company's content management platform would test these packages through experiments. The platform would measure responses, and compare the probability of a viewer clicking on

different potential packages for the same story.

The experiments were conducted on the homepage and article pages of Upworthy.com by randomly assigning different readers to see different headlines for the same story in recommendations to readers. The experiments allowed comparison between headlines, the number of participants that were shown a given package (impressions) and the number that clicked on the headline (clicks). Based on the experiment results, the editorial team would choose the winning headline, and from then it would be the only headline to be displayed on the homepage (Matias and Munger 2019).

At the core of Upworthy's A/B testing strategy was their belief that virality is a function of content and the clickability of the headline that displayed the content. However, cofounder and curator-in-chief Peter Koechley advocated that it was much easier to optimize on the headline - creating the headline that would attract the most clicks - than the content itself.[4] As such, Upworthy determined that headlines need to be optimized for different platforms. They argued that a headline works well for Google when everything readers want to know is presented to them within few words - meaning there is no need to click through. In contrast, a good social media headline seduces people to click through by creating a curiosity-gap. Pioneered by George Loewenstein's research (Loewenstein 1994), this method works by telling readers enough to whet their curiosity but not enough to fulfill it, which in turn drives them to click on the headline for wanting more of the story.

A/B testing titillating headlines without changing the underlying content was quite unique to the extent that it allowed the website to manage its con-

[4]https://www.slideshare.net/fmsignal/this-title-matters-more-than-my-talk-speaker-peter-koechley-upworthy

tent with insights obtained from experiments. This experimental approach indeed produced powerful results: Upworthy's repackaged videos and articles received an average of 75,000 likes per post on Facebook, about 12 times that of any other news organization, and the site spiked to 87 million unique viewers in December 2013 (Fitts 2014). Around the same time, Upworthy was being called the fastest-growing media company in the world and also won the fastest rising startup in TechCrunch's Crunchie awards (Kamenetz 2013).

3.2.3 Data

To study our research questions we leverage the Upworthy research archive (Matias and Munger 2019), which was made available to the public in April 2020 by academic researchers for research and educational purposes.[5] The complete archive consists of 32,488 experiments conducted at Upworthy.com from January, 2013 through April, 2015. For each experiment, the dataset includes viewer responses to each package in an experiment. A package refers to a pair of headline and the image that accompanies it. The dataset includes over 150 thousand packages with a median of 4 packages per experiment. Together, these packages received over 538 million impressions and over 8 million clicks. Each experiment included a median of 14,342 impressions and a median of 201 clicks per experiment.

Around half of the experiments in the sample compared both headlines and images. However, images are not available in the archive and only their unique identifier can be seen, therefore we only kept tests that used the same image

[5]Interested readers may find more information about the archive at https://upworthy.natematias.com/

and only varied the headlines. This left us with 16,645 valid tests and 72,787 headlines. The tests in our sample received over 260 million impressions and over 3.5 million clicks.

Figure 3.1 is an example of an experiment that shows the same story with four different headlines at Upworthy.com. Table 3.1 shows the summary of the key outcome variables across 121 weeks.

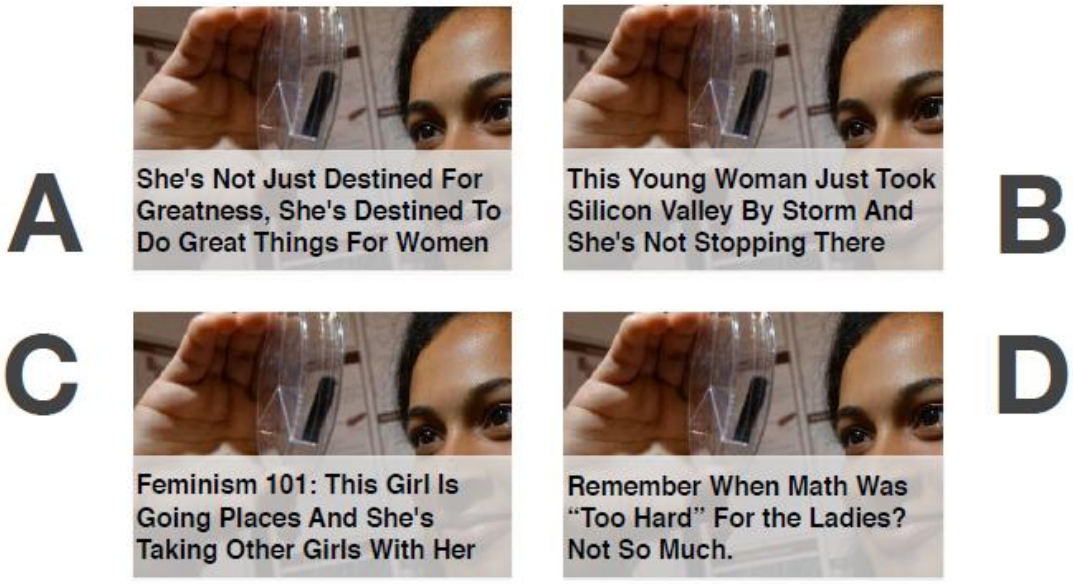

Figure 3.1: Example of Packages Tested at Upworthy

This dataset has a few unique features that make it a good match for our research purposes. First, it consists of a series of experiments, as such is free of selection-bias that usually poses a challenge using observational data. Second, the archive covers 121 weeks, thus has enough observations before and after the changes to the News Feed algorithm. Finally, the experiments conducted at the website are independent of Facebook, and thus the change of the News Feed algorithm bears no connection to the experiments or their outcomes. These

Table 3.1: Summary Statistics

	Mean	St.Dev.	Min	Max
No. of tests	137.56	70.63	5.00	323.00
No. of headlines	601.55	357.62	18.00	1617.00
No. of impressions	2,164,760.00	1,694,514.00	36,126.00	8,530,511.00
No. of. click	28,820.06	16,793.48	901.00	67,231.68

features altogether make this dataset conducive for studying the causal impact of the change to the News Feed algorithm at Facebook.

3.3 Research Design and Methodology

In this section, we describe the framework we adopted to study the impact of the algorithmic change in Facebooks' News Feed on both user and publisher behavior during our observation period.

3.3.1 Patterns in the Data and Descriptive Statistics

To begin with, in Figures 3.2 to 3.5 we show without employing any model how user and publisher behavior change over time, especially before and after the policy change by Facebook. A clear pattern emerges from the figures: the weekly number of tests, headlines per test increased throughout the observation

period, whereas the number of impressions per headline remained stable and clickthrough rates decreased.

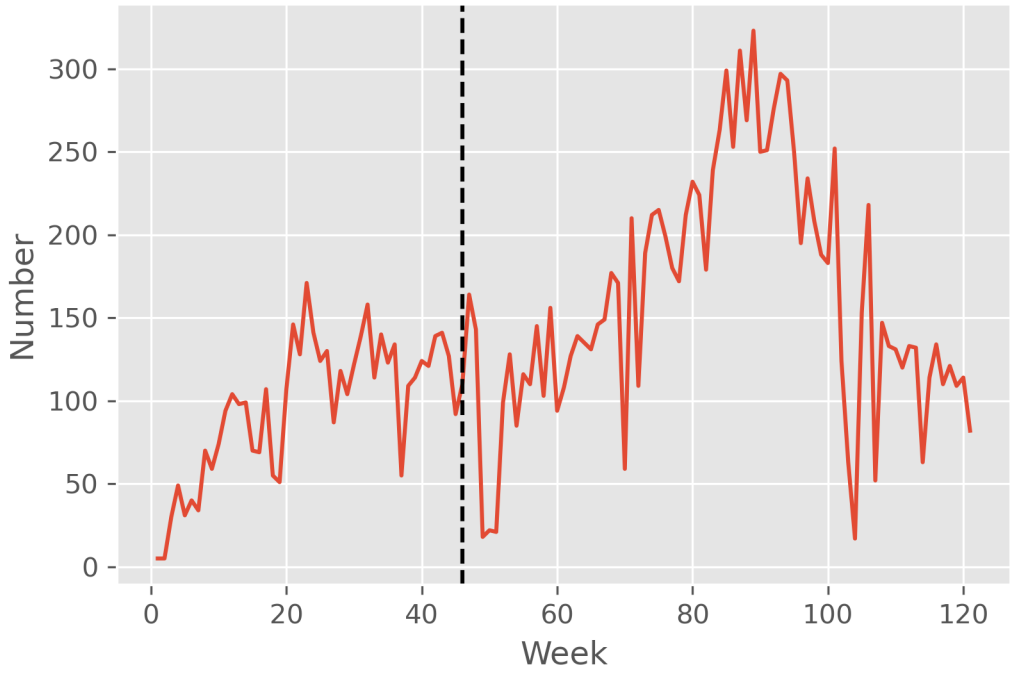

Figure 3.2: Number of Tests by Week

To test in a model-free manner whether the changes on our measures of interests were statistically and economically meaningful, we compared the means of the outcome measures before and after the algorithm change went into effect in week 47 of our observation period. Table 3.2 shows the results, which clearly suggest that the changes are significant both economically and statistically. Compared to the before period, in the period after the change went into effect, the publisher on average performed 65 more experiments per week and tested on average 0.73 more headline per experiment. The publisher also increased the number of impressions it assigned per test by 2000. On the user

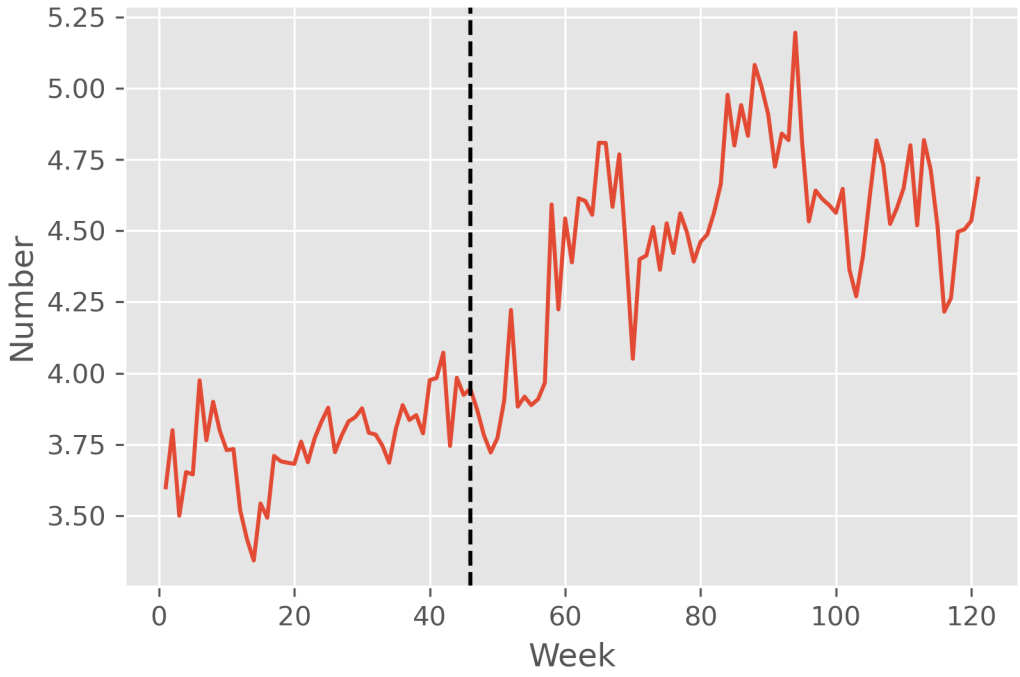

Figure 3.3: Number of Packages per Test by Week

side, we also see a significant change in behavior. The average clickthrough rate per experiment, in particular, decreased by about 0.8 percentage points. Considering that the average clickthrough rate was 2.1% before the change, this corresponds to almost a 40% decline in clickthrough rate.

The model-free analysis is a simple comparison of our measures of interest before and after the change went into effect. As such, it does not take into account differences that might result from editorial decisions. For example, headlines might have different features in some parts of the data period, which might play a role in users' decision clicking on them. Also, the data structure is nested, which implies that headlines within nests are not independent. Headlines were tested within different experiments, and experiments were conducted at dif-

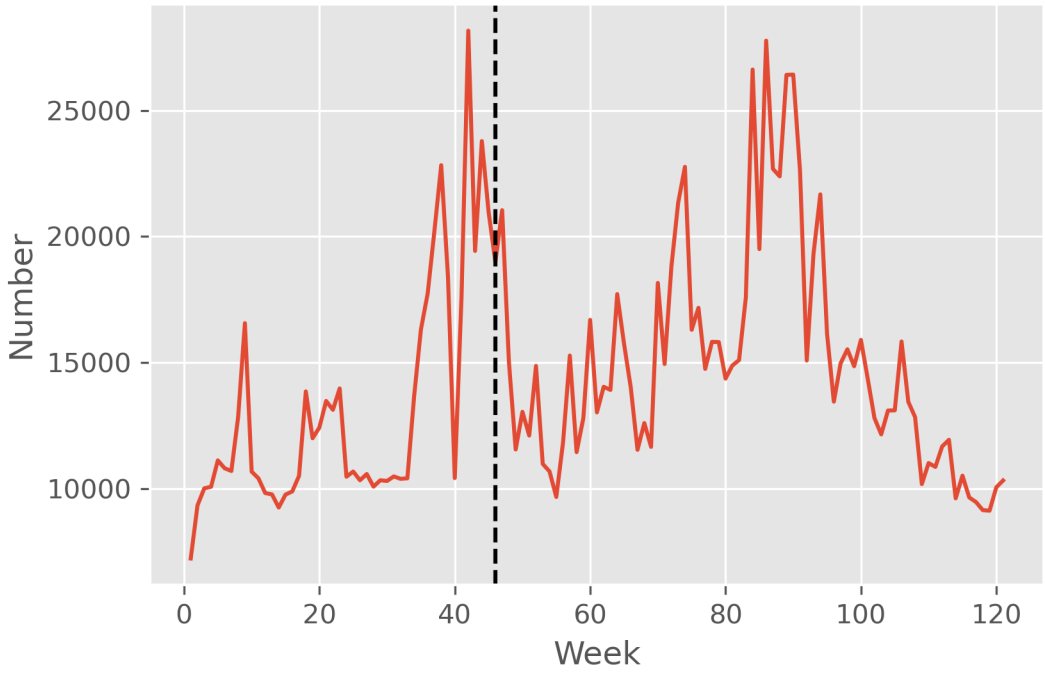

Figure 3.4: Number of Impressions per Test by Week

Table 3.2: Pre and Post Mean Differences

	Pre	Post	Difference
No. of tests	96.71	161.75	65.04
No. of headlines per test	3.76	4.49	0.73
No. of impressions per test	13135.34	15136.26	2000.92
Clickthrough rate per test	2.10%	1.30%	-0.80%

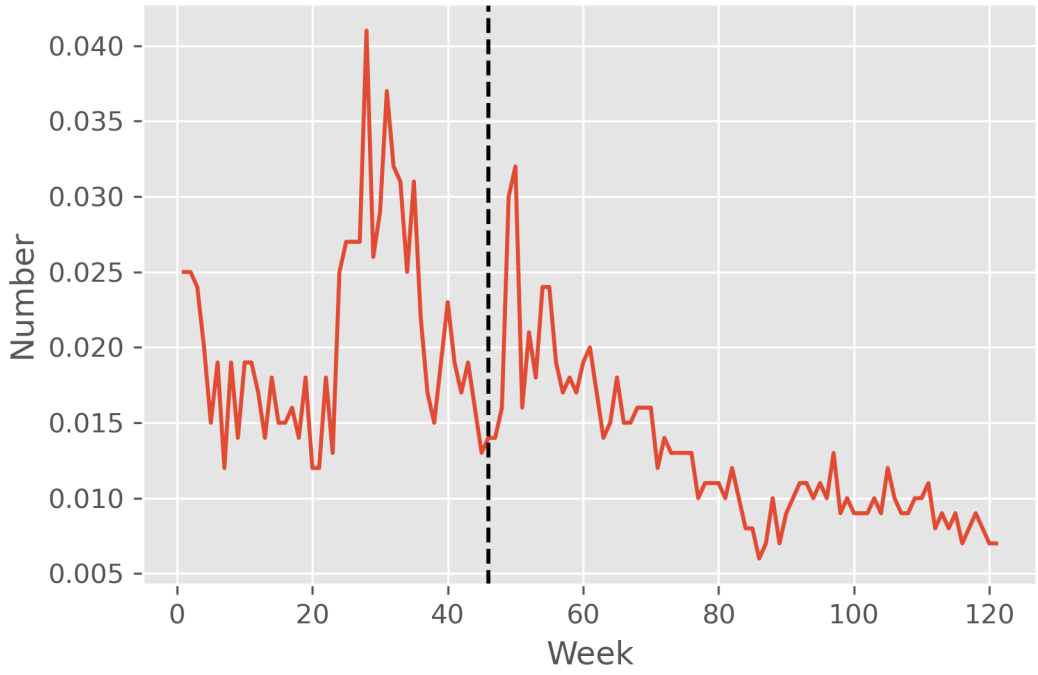

Figure 3.5: Clickthrough Rate per Test by Week

ferent weeks, which together cause dependencies that could be a factor in the changes we observe. Therefore, we need to build an empirical model that accounts for these factors and allows us to estimate the impact of the algorithm change net of all other factors. To do so, we first employ deep-learning to classify headlines into clickbait or non-clickbait classes.

3.3.2 Clickbait Classification

It is critical that our analysis differentiates between clickbait and non-clickbait headlines for several reasons. First, clickbaits are designed to lure users into clicking by stimulating their curiosity, therefore we expect them to generate

higher clickthrough rates. Second, the algorithmic change might increase users' awareness towards clickbaits, thus it might impact their attitude towards them differently after the change went into effect compared to non-clickbaits.

We also know from our descriptive analysis that the number of headlines per week increased at a faster pace in the post period. Since, clickbaits are expected to generate higher clicks, not controlling for them in our analysis will cause our estimates to be under or overestimated depending on whether the number of clickbait headlines followed a similar or different path then the number of headlines per week.

To create a supervised classification model we employed deep-learning that does not require feature engineering. Using deep-learning rather than feature engineering eliminates the need to manually specify the features in headlines that might play a role in the classification task and delegates this process to an automated deep-learning architecture. Furthermore, we used pre-trained word embeddings to transform the words in the corpus to embeddings. The embeddings have 300 dimensions and were trained using a distributed sub-word embedding technique on 1.67 million Facebook posts that were shared by a set of mainstream and unreliable media within January 1st, 2014 – December 31st, 2016. The technique was introduced by Rony, Hassan, and Yousuf (2017), uses an extension of the continuous skip-gram model which takes into account subword (substring of a word) information and is called as Skip-Gram. The hallmark of the method is that it breaks down words into subwords, rather than treating each word as a unit, and aims to correctly predict the context subwords of a given subword. This extension allows sharing the representations across words, thus allowing to learn reliable representations for rare words. Using

neural network, the method learns the mapping between the output and the input. The weights to the hidden layer in the neural network form the vector representations of the subwords.

Rony, Hassan, and Yousuf (2017) used headlines and bodies from 1.67 million Facebook posts to learn 300 dimensional word embeddings using this model. It allows having richer word embeddings which capture the details of semantic, conceptual and contextual information. Compared to Google News dataset's 100 billion embeddings the corpus has 477,236 unique embeddings. The authors demonstrate that even though the size of their corpus is significantly smaller than the Google News dataset, it contributes significantly more to the clickbait classification task. According to the authors, this observation can be explained through the fact that the embeddings learned from their corpus having more domain specific knowledge than the Google News dataset. In light of this, we decided to use these pre-trained word embeddings in our clickbait classification task.

We use these embeddings to map headlines to a vector space and then employ a deep recurrent neural network (RNN) architecture with a gated recurrent unit layer (GRU). In the last few years, there have been incredible success applying RNNs to a variety of problems: speech recognition, language modeling, translation, and image captioning. They have also proven to be very effective in text classification problems (Anand, T. Chakraborty, and N. Park 2017). Compared to mainstream ML methods, deep-learning based methods, especially RNNs, can capture the dependencies between the beginning and end of a sentence, hence classify texts more accurately. For example, traditional ML methods would incorrectly classify: "This place lacks good food, good service,

good hygiene." as possessing positive sentiment because the prevalence of the word "good".

The deep network takes the sequence of embedding vectors as input and converts them to a compressed representation. The compressed representation effectively captures all the information in the sequence of words in the text. A very common problem with RNN based networks is the tendency to overfit. To overcome this problem we added a dropout feature to the network. The fully connected layer takes the deep representation from the RNN/GRU and transforms it into the final output class scores. This component is comprised of fully connected layers along with batch normalization for regularization. Finally in the ouput layer the softmax function is applied as a classifier. Figure 3.6 concisely shows the architecture of our deep-learning classifier.

We trained this classifier on a dataset with pre-labeled headlines. This dataset has been utilized by researchers in computer science, It was curated by A. Chakraborty et al. (2016) and was also used by Rony, Hassan, and Yousuf (2017). It contains 32, 000 headlines of news articles which appeared in 'WikiNews', 'New York Times', 'The Guardian', 'The Hindu', 'BuzzFeed', 'Upworthy', 'ViralNova', 'Thatscoop', 'Scoopwhoop', and 'ViralStories'. Each of these headlines were manually labeled either as a clickbait or a non-clickbait by at least three researchers. There are 15,999 clickbait headlines and 16,001 non-clickbait headlines in this dataset. We used this labeled dataset to develop and train our automatic clickbait classification model.

Our best performing model achieved 98% accuracy on a labeled validation set after being trained on a labeled training set. We used this model to classify the 72,787 headlines in our Upworthy archive. The model classified 92% of the

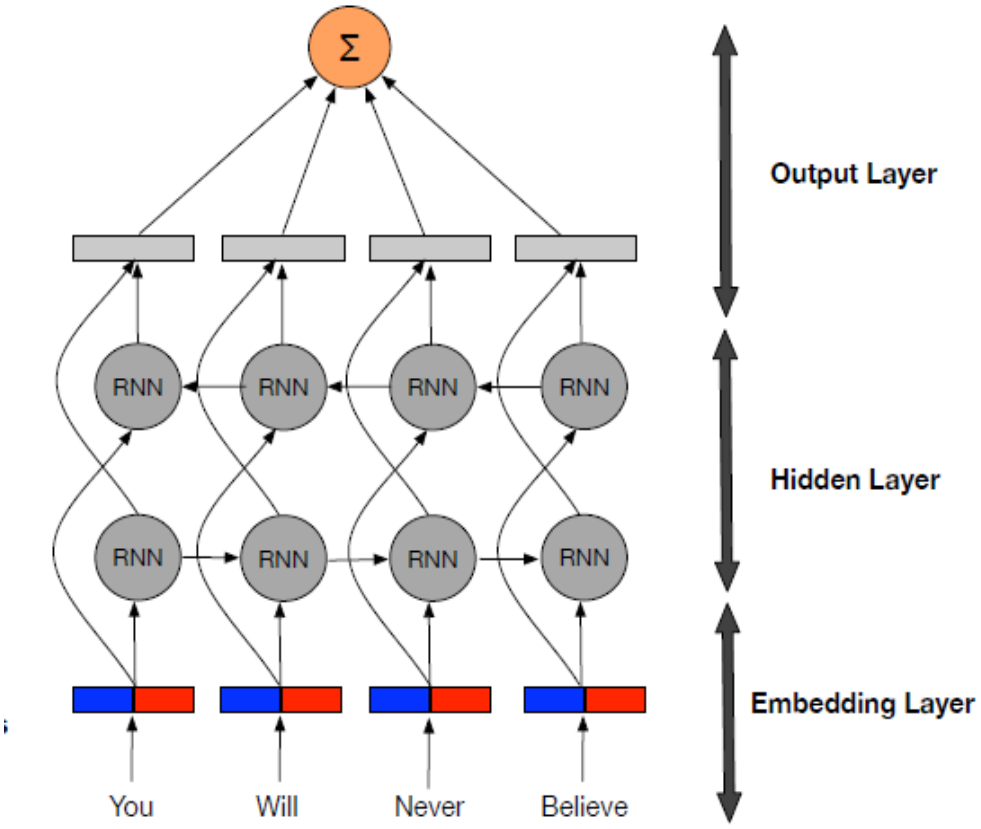

Figure 3.6: Deep Learning Architecture for Headline Classification

headlines as clickbait. Figure 3.7 shows how the number of clickbait headlines per test changed on a weekly basis. Similar to the number of headlines, the number of clickbait headlines per test increased throughout the data period.

3.3.3 Empirical Model

Our model-free analysis has revealed that clickthrough rates declined significantly in the weeks after Facebook announced its change to its News Feed algorithm. Now, we aim to construct a model that can allow us statistically measure and answer several questions: First, we seek to find whether the policy change

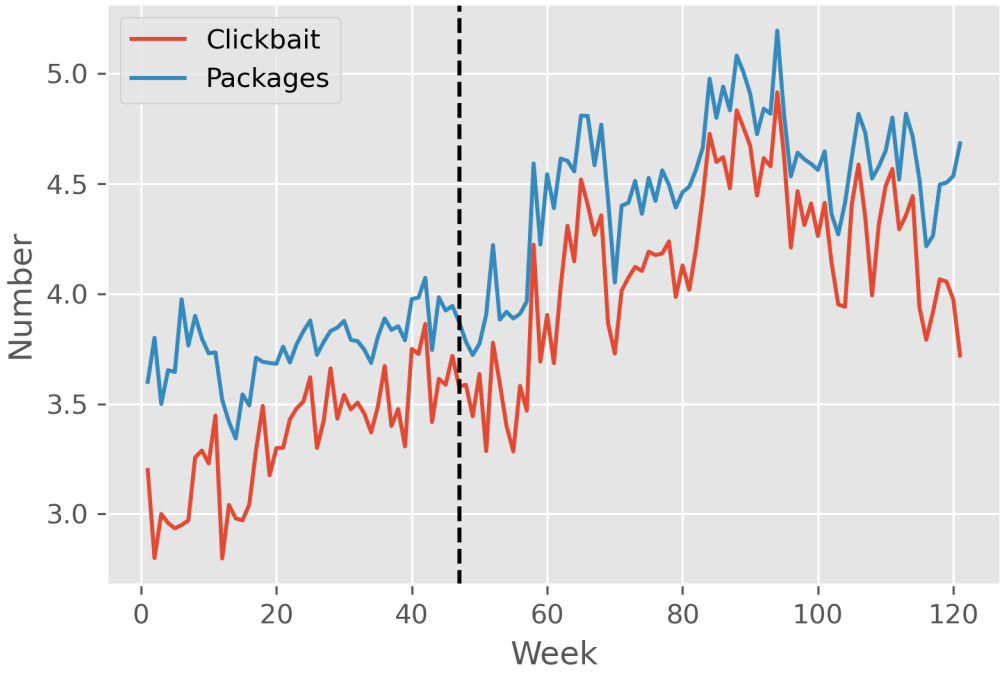

Figure 3.7: Number of Cliclbait and Packages per Test by Week

had any impact on clickthrough rates. Second, we want to understand the effect of clickbait on clicks. Finally, we seek to explore whether the policy change impacted clicks differently for clickbait and non-clickbait headlines.

Because Facebook's purpose in changing the algorithm was aimed at reducing the presence of low-quality posts and eliminating clickbaits on Facebook, we might expect that clicks on clickbait headlines suffered more than non-clickbait ones in the period after the change. This would be the case if the change to the algorithm increased awareness towards clickbait and caused a change in users' preferences. Alternatively, users might not have distinguished between headlines and both types might have suffered at similar amounts. This would apply to the case in which the change in the algorithm did not drive users' preferences.

As such, the question is subject to competing hypotheses and ultimately is an empirical one.

To answer our questions we build a statistical model that exploits the nested structure of the data. We take each experiment as the unit of analysis, which allows us to model the dependency of the headlines within experiments. Our identification strategy leverages the fact that Facebook's decision to change the News Feed algorithm is an exogenous event that occurred at a specific date, and is independent from the experiments being employed at Upworthy.com. We therefore, employ pre-post estimation to identify the causal impact of the algorithm change. Our model has the following structure:

$$\mathrm{y}_i = \alpha + \beta_1 Post_i + \beta_2 Clickbait_i + \beta_3 Post_i \times Clickbait_i + \epsilon_i \tag{3.1}$$

where y_i is the average clickthrough rate for packages in test i, defined as the ratio of the total number of number clicks to the total of impressions that all the headlines in experiment i received. The indicator variable $Post_i$ is 1 if test i occurred after the algorithm change and 0 otherwise. $Clickbait_i$ represents the fraction of clickbait headlines in test i, and ϵ_i is the error term. We cluster the error term at the week level to account for clustering effects that might arise from specific weekly trends.

By allowing the coefficients of $Clickbait_i$ to vary in the periods before and after the policy change, we are able to investigate whether the policy change impacted headlines differently using this model. Specifically, if β_3 is estimated to be signifcantly different than 0, then we can conclude that clickbaits are impacted differently than non-clickbait headlines due to the policy change. If,

however, β_3 is estimated to be not signifcantly different than 0, we would conclude that users did not differentiate between clickbaits and non-clickbaits, and that clicks decreased similarly in both types.

3.4 Findings

Before we discuss our findings regarding the impact of the News Feed algorithm change on users' behavior at Upworthy.com from our empirical models, we reiterate what our analysis has revealed about the impact on the publisher. As we have shown in table 3.2, compared to the pre-change period the publisher performed on average 65 more experiments and tested 0.73 more headline per experiment in the post-change period. The publisher also increased the number of impressions it assigned to the experiments in the post-period by 2000. Regarding the classification of the headlines, our analysis has also revealed that the publisher produced clickbaits at a higher rate after the change went into effect.

Moving to user behavior, we start with discussing the findings from our model, which we estimate using maximum likelihood without headline class first and then include clickbait to evaluate the sensitivity of our estimates. We show our estimation results in table 3.3. Columns (1) and (2) in the table show the results without and with headline classification, respectively. Across both specifications the impact of the algorithm change is both economically and statistically significant. Specifically, we find that compared to the period before the change, clickthrough rates in the post-period declined on average by about 1 percentage points. Recall from table 3.2 that the average clickthrough rate

before the change was 2.1%. In light of this, the impact of the change on click-through rates is about a 50% decline, which is quite substantial.

Table 3.3: Estimation Results

Post	-0.0098***	-0.0099***	-0.0074***
	(0.0010)	(0.0010)	(0.0020)
Clickbait		0.0066***	0.0085***
		(0.0010)	(0.0020)
Clickbait × Post			-0.0028
			(0.0020)
No. of observations	16,645	16,645	16,645

Notes:$***p < 0.001$, $**p < 0.01$, $*p < 0.05$. Standard errors are clustered at the week level and shown in parentheses.

Consistent with our intuition and the publisher's publication strategy, we find that clickbaits in general generate larger clickthrough rates. Our results in column (2) of the table reveal that tests that contain only clickbait headlines increase clickthrough rates on average by 0.66 percentage points compared to tests that have only non-clickbaits. In column (3) we show the results for the full model, which includes time-varying coefficients of clickbaits. We see that the estimate of β_3 is not statistically significant. This suggests that clickthrough rates did not decrease differently for clickbaits after the change to the News Feed algorithm went into effect. The results in column (3) also suggest that, adding time-varying coefficients into the model does not change the sign of $Post_i$ and has a minmial effect on its size. This shows that clickthrough rates on average still decreased in the post-change period even after allowing headline features to have time-varying effects on clickthrough rates.

3.5 Discussion and Suggested Mechanism

What is the possible explanation for the significant decrease in clickthrough rates for the headlines Upworthy tested on its website after the change in the News Feed algorithm? One potential explanation deals with the possibility that Upworthy made an editorial decision to change the features of its headlines after the algorithmic change at Facebook, which in turn might have caused a shift in user behavior. In particular, Upworthy might have chosen to produce fewer clickbait and more real headlines. However, as we have shown in figure 3.7, clickbait proportions in experiments became even more prevalent at Upworthy after the update to the News Feed algorithm. Essentially, Upworhty continued to produce more and not fewer clickbaits and headlines after the algorithm change went into effect. In the absence of a deliberate editorial change on the feature of the headlines, we have little reason to expect a change in users' choices.

Another potential explanation deals with the possibility that the update to the News Feed algorithm caused a change in user preference through increased awareness to clickbait headlines. If this was a likely mechanism then we would possibly observe a differential impact on clickthrough rates for clickbait and non-clickbait headlines. Recall, however, that the results from our time-varying analysis does not support this explanation. We found that the changes in clickthrough rates were not statistically different between clickbait and non-clickbait headlines. This fails to provide evidence that the change in the News Feed algorithm impacted clickthrough rates by changing customers' preferences, instead it must have impacted them by changing their behavior.

In light of this, we turn our attention to another potential explanation, which suggests a selection mechanism. As we have discussed earlier, soon after its inception, online traffic to Upworthy.com went on an upward trajectory and it peaked in December 2013. This peak in traffic was so high that it was accompanied by attention of the media, which resulted in Upworthy being awarded numerous awards including the recognition of the fastest-growing media company in the world (Matias and Munger 2019).

Arguably, the main reason Upworthy.com gained substantial traffic was because of the referral traffic it had been receiving from Facebook. Upworthy had been meticulously creating appealing shareable content for the social platform since it began to publish. In fact, before the algorithmic change to News Feed, Upworhty content was being shared on Facebook nearly eight times the rate of the next comparable site (Abebe 2014). With shares came traffic, because every click on a shared post from Upworthy was taking customers to the website.

We suggest that the act of clicking on Upworthy posts at Facebook was serving as a selection process for identifying people who had a higher preference for Upworthy content. After clicking on Upworthy content at Facebook, users would arrive at the website and it is unlikely that they would return to Facebook quickly once they consumed the content they clicked on. Instead, it is more likely that they would spend time on looking and possibly clicking on at least several other headlines at the website.

However, the change to Facebook's News Feed algorithm, caused a disruption in this selection mechanism. The News Feed algorithm started to demote Upworthy content, which resulted in less visibility for Upworthy content shared by Facebook users. As shown in previous research, the order in which a post is

presented to users in their News Feed significantly changes the probability of it being clicked by them (Bakshy, Messing, and Adamic 2015). As such, demoting Upworthy content in the News Feed resulted in less visibility, which in turn resulted in fewer clicks and referral traffic to Upworthy.com.

In figures 3.8 to 3.10 we show how number of weekly users, pageviews and average seesion duration at Upworthy.com changed. Consistent with our explanation, we see all metrics had been on an upward trend before the algorithm change, but started to steeply decline after the algorithm change went into effect.

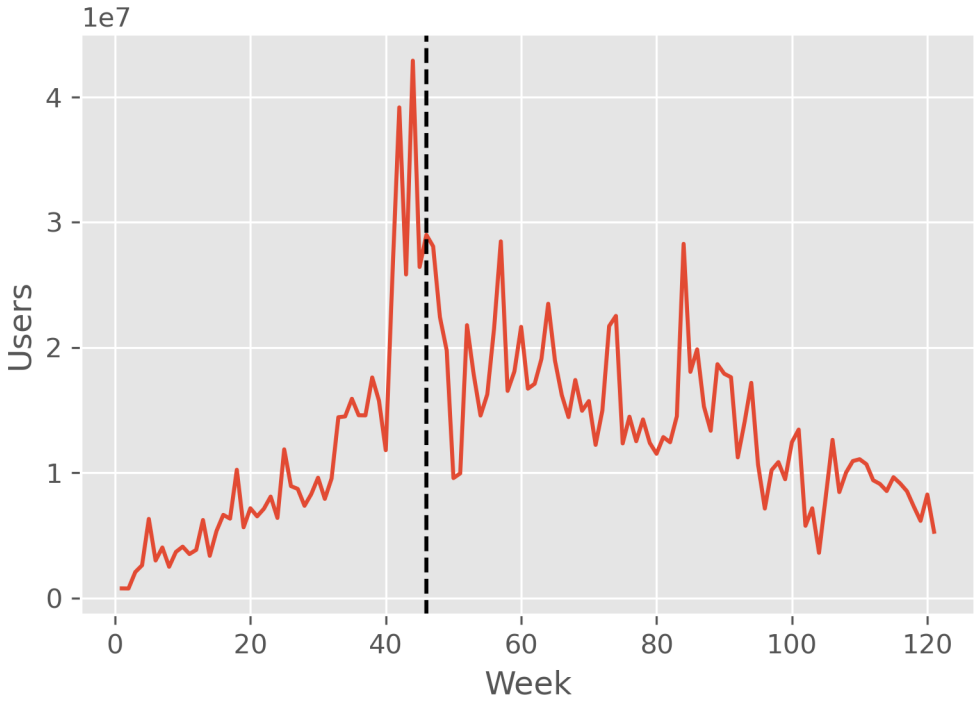

Figure 3.8: Number of Users by Week

Because referral traffic to Upworthy website dropped significantly after the change to the News Feed algorithm, the composition of the visitors was

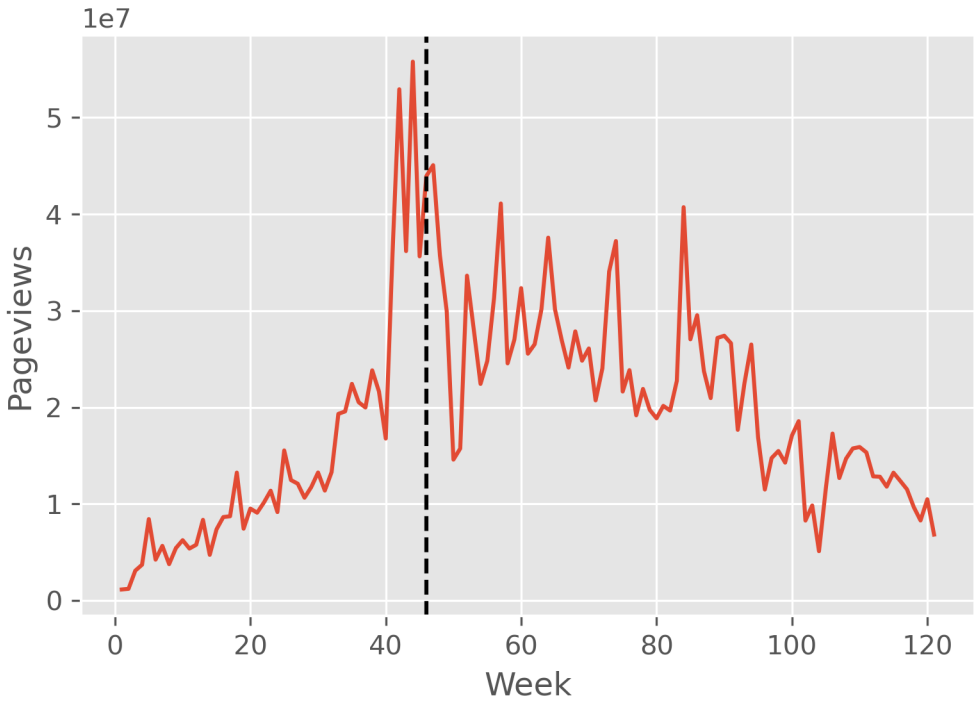

Figure 3.9: Number of Pageviews by Week

changed. Although, the company continued to assign similar number of impressions per headline in its weekly experiments, the visitors receiving the impressions were more likely to be organic visitors who already had been following the website and were coming directly to the website rather than being referred by Facebook. As a result of this change in visitor composition, Upworthy received fewer clicks from similar number of impressions in their online experiments.

To test this suggested mechanism, ideally we would have access to Upworthy.com referral traffic data from Facebook. With this data we could compare the changes in traffic to Upworthy website coming from Facebook before and after the algorithm change. Unfortunately, we do not have access to this data, instead

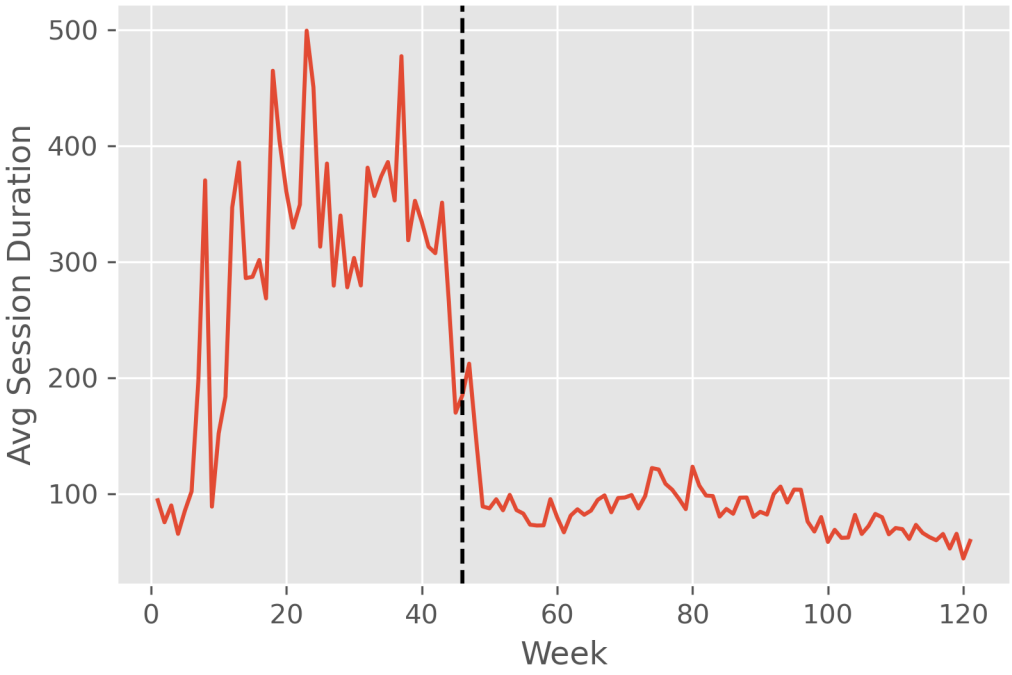

Figure 3.10: Average Session Duration by Week

we resort to Google trend data and leverage search data for keyword "Upworthy" as a proxy for referral traffic from Facebook. If the algorithm change had indeed impacted traffic to Upworthy website, we should observe a decline in Google search trend for "Upworthy" around the time when the change took effect.

Furthermore, we compare "Upworthy" search trend to that of "Buzffeed", another digital media company. Buzzfeed makes a good comparison because its publishing strategy was similar to that of Upworthy during our observation period. The company was well known for its visually appealing and shareable content on social media platforms. Similar to Upworthy, the company had gained popularity through its clickbait headlines. In fact, multiple studies in

computer science leveraged Buzzfeed headlines for training their clickbait classification models (Eidnes 2015; Thakur 2016; A. Chakraborty et al. 2016; Rony, Hassan, and Yousuf 2017).

However, Buzzfeed also differed from Upworthy on several fronts. First, in addition to content aggregation, the company also created and published original content on social platforms. Second, the company was reported to be a close partner of Facebook, because it had been purchasing ads on the platform (Carlson 2014). In light of this, it is likely in the eyes of Facebook executives, Buzzfeed content had been flagged as high quality and therefore News Feed-worthy compared to Upworthy content.

Figure 3.11 shows how search volume for "Upworthy" and "Buzzfeed" at Google changed throughout January 2013 to April 2015. As the respective lines on the figure shows, search volume increased steadily for both keywords until December 2013 and started to decline sharply afterwards for "Upworthy", whereas it continued to increase for "Buzzfeed". The declining trend for "Upworthy" was persistent after the change went into effect in December 2013 and seems to have stabilized towards the end of our observation period.

To formally evaluate the impact of the algorithm change on the Google search trend changes, we fit the following differences-in-differences model to the search volume data:

$$y_{it} = \lambda + \theta_1 Post_{it} + \theta_2 Upworthy_{it} + \theta_3 Post_{it} \times Upworthy_{it} + \eta_{it} \tag{3.2}$$

where y_{it} is the search volume for keyword i in week t at Google. The indicator variable $Post_{it}$ is 1 if week t occurred after the algorithm change and 0 otherwise.

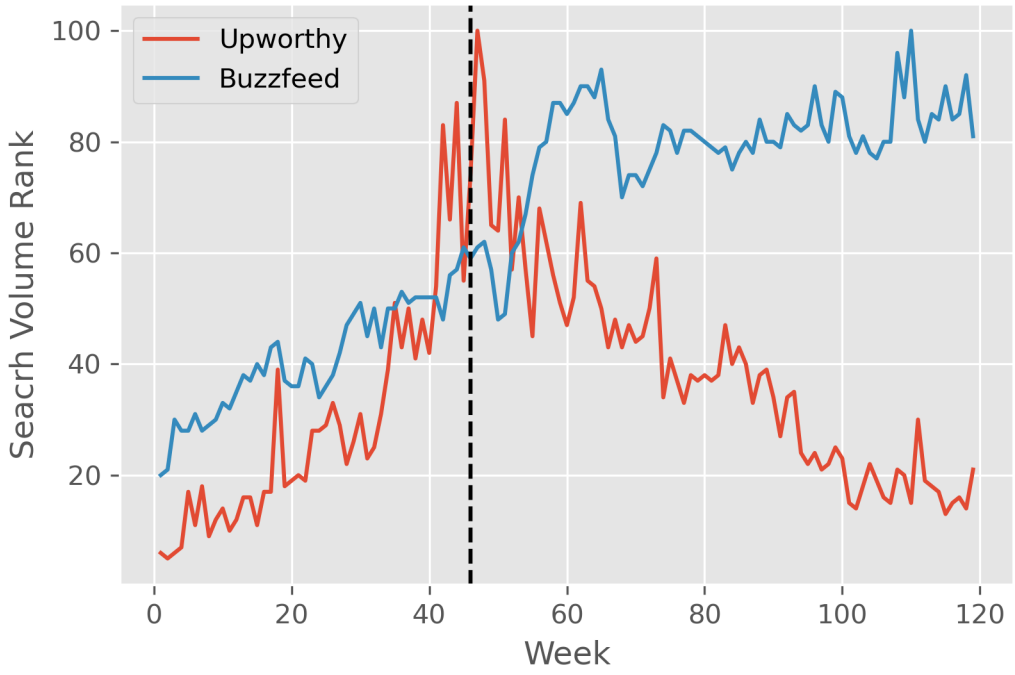

Figure 3.11: Google Search Trends by Week

$Upworthy_{it}$ is an indicator that is 1 if the keyword i in week t was "Upworthy" and 0 otherwise, and η_{it} is the error term. Our focus in this model is on θ_3, which captures the difference-in-differences in search volume. If θ_3 is estimated to be signifcantly different than 0, then search volume for "Upworthy" was impacted differently than "Buzzfeed" due to the policy change.

The identification assumption behind this analysis is the parallel-trends assumption (Angrist and Pischke 2008), which states that in the absence of the algorithm change, search volume for both keywords would have changed similarly from the period before the change to the post-period. Even though it is impossible to verify this assumption, the trends in the period before the change went into effect can be used to gauge whether it is plausible or not. As shown in

3.11 the trends for both keywords follow a similar upward trend until the week of the policy change, hence we can conclude that the parallel-trends assumption is credible in our context.

Table 3.4 shows the estimation results. We find the impact of the algorithm change on "Upworthy" search volume was significantly negative compared to "Buzzfeed" search volume on Google. After the change went into effect, "Upworthy" search volume declined by 29 points compared to "Buzzfeed" search volume. This is evidence that the algorithm change at News Feed had been designed to demote content from specific publishers, such as Upworthy, which is consistent with what was reported in the media (Kafka 2013). Moreover, it provides support that our findings regarding the drop in clickthrough rates at Upworthy.com after the algorithmic change was primarily driven by the decline in referral traffic the website has been receiving from Facebook. In light of this, we conclude that the change in News Feed algorithm decreased traffic to the website by demoting Upworthy posts, thereby decreasing visibility and likelihood of clicks.

3.6 Conclusions

As society becomes increasingly reliant to platforms for its daily interactions, the algorithms employed in these platforms for facilitating them become more powerful. Across many platforms their status have been elavated from simple computational tools to gatekeepers. They determine what customers are being exposed to, thus shape visibility, sharing and flow of information (Tufekci 2015).

Recent research in marketing and social sciences have documented the im-

Table 3.4: Google Trends Difference-in-Differences Estimation Results

Post	38.34***
	(1.86)
Upworthy	-11.87***
	(3.35)
Upworthy × Post	-29.24***
	(4.19)
No. of observations	242

Notes: $***p < 0.001$, $**p < 0.01$, $*p < 0.05$. Robust standard errros shown in parentheses.

pacts algorithms can have on customer behavior and feelings within a platform or website (e.g., Hauser et al. 2009; Kramer, Guillory, and Hancock 2014; Bakshy, Messing, and Adamic 2015). However, the impacts of algorithms on customer behavior can reach beyond the confines of the platform in which they are employed, and research into this potential effect has been missing.

In this research we aim to fill this gap by studying the impact of a platform's decision to change its ranking algorithm by flagging content from certain publishers as low-quality and demoting them had on users' beahvior towards content of a publisher that operates within the platform but also runs its own website. We focus specifically on the largest online social media platform: Facebook, and investigate how the first publicly announced major update to its News Feed algorithm, which occurred on December 2, 2013, impacted users' behavior at Upworthy.com.

Utilizing the Upworthy research archive, an experimental dataset covering

a period of 121 weeks (January 2013 - April 2015), we employ deep-learning techniques to classify headlines and develop an empirical model that exploits the nested structure of the data to estimate the causal effect of the change. Our identification relies on the fact that the algorithm change was an exogenous event with respect to the experiments performed at Upworthy.com.

Our findings show that the change to the News Feed algorithm caused a significant decline in clickthrough rates for headlines Upworthy was testing on its website. To uncover the underlying mechanism for the observed changes, we leverage Google trend data as a proxy for referral traffic to the Upworhty website. We analyze how search volume for the keywords "Upworthy" and "Buzzfeed" varied throughout our observation period. Our analysis indicates that traffic to the website declined significantly in the post-change period, suggesting that changing the News Feed algorithm reduced referral traffic Upworthy had been receiving from Facebook. We posit that this change disrupted the selection mechanism of users who had a higher preference for Upworthy posts shared by other users, which eventally led to to a sharp decline in clickthrough rates at Upworthy.com.

Our results offer important managerial implications. First, this research presents evidence that the impact algorithms have on users can reach beyond the platform they are employed. As such, we recommend managers to closely monitor the announcememts of the platforms they are operating in, and keep abreast of the changes they make to their algorithms. Second, our findings suggest that changes to platforms' algorithms can alter the composition of the users visiting the business' website. Therefore, we emphasize the need for managers to be cognizant of this phenomenon and establish proactive measures to hedge

against sudden changes the platform can make to its algorithm. Doing business in the same way as if nothing had changed is simply not effective and likely to waste useful resources.

As with all empirical research our study has a number of limitations, which should be acknowledged and perhaps addressed in future research. First, our study focused on a single publisher in a specific platform. Therefore, replication across other platforms would be needed to build empirical generalizations on this topic. Second, our dataset does not allow us to conduct a more granular analysis at the user level. Without data on individual user behavior we are unable to rule out other possible mechanisms that might explain our findings. Finally, our context does not allow us to study what would be the ideal response of the publisher in response to the announced changes from the platform. We believe this is a fruitful topic for future research.

References

Agarwal, Ashish, Kartik Hosanagar, and Michael D Smith (2011). "Location, location, location: An analysis of profitability of position in online advertising markets". In: *Journal of Marketing Research* 48.6, pp. 1057–1073.

Anand, Ankesh, Tanmoy Chakraborty, and Noseong Park (2017). "We used neural networks to detect clickbaits: You won't believe what happened next!" In: *European Conference on Information Retrieval*. Springer, pp. 541–547.

Angrist, Joshua D and Jörn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, NJ.

Ascarza, Eva (2018). "Retention futility: Targeting high-risk customers might be ineffective". In: *Journal of Marketing Research*, 55(1):80–98.

Bakshy, Eytan, Solomon Messing, and Lada A Adamic (2015). "Exposure to ideologically diverse news and opinion on Facebook". In: *Science* 348.6239, pp. 1130–1132.

Carlson, Nicholas (2014). "Buzzfeed CEO: Here's Why Facebook Isn't Crushing Us." In: URL: `https://www.businessinsider.com/buzzfeed-traffic-2014-2`..

Chakraborty, Abhijnan et al. (2016). "Stop clickbait: Detecting and preventing clickbaits in online news media". In: *2016 ieee/acm international conference on advances in social networks analysis and mining (asonam)*. IEEE, pp. 9–16.

Chan, Tat Y and Young-Hoon Park (2015). "Consumer search activities and the value of ad positions in sponsored search advertising". In: *Marketing Science* 34.4, pp. 606–623.

Chen, Le, Alan Mislove, and Christo Wilson (2016). "An empirical analysis of algorithmic pricing on amazon marketplace". In: *Proceedings of the 25th International Conference on World Wide Web*, pp. 1339–1349.

Chung, Tuck Siong, Roland T Rust, and Michel Wedel (2009). "My mobile music: An adaptive personalization system for digital audio players". In: *Marketing Science* 28.1, pp. 52–68.

Fitts, Alexis Sobel (2014). "The king of content. how Upworthy aims to alter the Web, and could end up altering the world." In: *Columbia Journalism Review.* URL: https://archives.cjr.org/feature/the%5C_king%5C_of%5C_content.php.

Kamenetz, A (2013). "How upworthy used emotional data to become the fastest growing media site of all time." In: URL: https://www.fastcompany.com/3012649/how-upworthy-used-%20emotional-data-to-become-the-fastest-growing-media-site-of-all-time..

Kacholia, Varun and Minwen Ji (2013). "Helping You Find More News to Talk About." In: URL: https://about.fb.com/news/2013/12/news-feed-fyi-helping-you-find-more-news-to-talk-about/..

Fong, Nathan et al. (2019). "Targeted Promotions on an E-Book Platform: Crowding Out, Heterogeneity, and Opportunity Costs". In: *Journal of Marketing Research*, 56(2):310–323.

Ghose, Anindya, Panagiotis G Ipeirotis, and Beibei Li (2014). "Examining the impact of ranking on consumer behavior and search engine revenue". In: *Management Science* 60.7, pp. 1632–1654.

Hauser, John R et al. (2009). "Website morphing". In: *Marketing Science* 28.2, pp. 202–223.

Jerath, Kinshuk et al. (2011). "A "position paradox" in sponsored search auctions". In: *Marketing Science* 30.4, pp. 612–627.

Kafka, Peter (2013). "Like This if You Like Pandas! Facebook Says Publishers Shouldn't Fret About News Feed Changes." In: URL: http://allthingsd.com/20131206/like-this-if-you-like-pandas-facebook-says-publishers-shouldnt-fret-about-news-feed-changes/..

Swisher, Kara (2011). "Ahead of Earnings Next Week, Demand Media Shares Drastic Dip Due to Googley Panda-Monium." In: URL: http://allthingsd.com/20110427/demand-shares-drastic-dip-due-to-googley-panda-monium/..

Kramer, Adam DI, Jamie E Guillory, and Jeffrey T Hancock (2014). "Experimental evidence of massive-scale emotional contagion through social networks". In: *Proceedings of the National Academy of Sciences* 111.24, pp. 8788–8790.

Eidnes, Lars (2015). "Auto-Generating Clickbait With Recurrent Neural Networks". In: URL: https://larseidnes.com/2015/10/13/auto-generating-clickbait-with-recurrent-neural-networks/..

Loewenstein, George (1994). "The psychology of curiosity: A review and reinterpretation." In: *Psychological bulletin* 116.1, p. 75.

Matias, J Nathan and Kevin Munger (2019). "The Upworthy Research Archive: A Time Series of 32,488 Experiments in US Advocacy." In: *MIT Code*.

Abebe, Nitsuh (2014). "Watching Team Upworthy Work Is Enough to Make You a Cynic. Or Lose Your Cynicism. Or Both. Or Neither." In: *New York Magazine.* URL: https://nymag.com/intelligencer/2014/03/upworthy-team-explains-its-success.

Rony, Md Main Uddin, Naeemul Hassan, and Mohammad Yousuf (2017). "Diving deep into clickbaits: Who use them to what extents in which topics with what effects?" In: *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pp. 232–239.

Rutz, Oliver J and Michael Trusov (2011). "Zooming in on paid search ads—A consumer-level model calibrated on aggregated data". In: *Marketing Science* 30.5, pp. 789–800.

Wong, Danny (2015). "In Q4, Social Media Drove 31.24% of Overall Traffic to Sites." In: URL: https://www.shareaholic.com/blog/social-media-traffic-trends-01-2015/..

Thakur, Abhishek (2016). "Identifying Clickbaits Using Machine Learning." In: URL: https://www.linkedin.com/pulse/identifying-clickbaits-using-machine-learning-abhishek-thakur..

Tufekci, Zeynep (2015). "Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency". In: *Colo. Tech. LJ* 13, p. 203.