Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Design and Evaluation of Optimal Free Trials

Hema Yoganarasimhan, Ebrahim Barzegary, Abhishek Pani

To cite this article:
Hema Yoganarasimhan, Ebrahim Barzegary, Abhishek Pani (2023) Design and Evaluation of Optimal Free Trials. Management Science 69(6):3220-3240. https://doi.org/10.1287/mnsc.2022.4507

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 69, No. 6, June 2023, pp. 3220–3240
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Design and Evaluation of Optimal Free Trials

Hema Yoganarasimhan,[a,*] Ebrahim Barzegary,[a] Abhishek Pani[b]

[a] Foster School of Business, University of Washington, Seattle, Washington 98195; [b] Bright Machines, San Francisco, California 94107
*Corresponding author

Contact: hemay@uw.edu, https://orcid.org/0000-0003-0703-5196 (HY); ebar@uw.edu, https://orcid.org/0000-0002-7268-8778 (EB); abhishek.pani@gmail.com (AP)

Received: June 2, 2020
Revised: May 1, 2021, November 16, 2021
Accepted: January 10, 2022
Published Online in Articles in Advance: August 10, 2022

https://doi.org/10.1287/mnsc.2022.4507

Abstract. Free trial promotions are a commonly used customer acquisition strategy in the Software as a Service industry. We use data from a large-scale field experiment to study the effect of trial length on customer-level outcomes. We find that, on average, shorter trial lengths (surprisingly) maximize customer acquisition, retention, and profitability. Next, we examine the mechanism through which trial length affects conversions and rule out the demand cannibalization theory, find support for the consumer learning hypothesis, and show that long stretches of inactivity at the end of the trial are associated with lower conversions. We then develop a personalized targeting policy that allocates the optimal treatment to each user based on individual-level predictions of the outcome of interest (e.g., subscriptions) using a lasso model. We evaluate this policy using the inverse propensity score reward estimator and show that it leads to 6.8% improvement in subscription compared with a uniform 30-days for-all policy. It also performs well on long-term customer retention and revenues in our setting. Further analysis of this policy suggests that skilled and experienced users are more likely to benefit from longer trials, whereas beginners are more responsive to shorter trials. Finally, we show that personalized policies do not always outperform uniform policies, and we should be careful when designing and evaluating personalized policies. In our setting, personalized policies based on other methods (e.g., causal forests, random forests) perform worse than a simple uniform policy that assigns a short trial length to all users.

History: Accepted by Duncan Simester, marketing.
Supplemental Material: The data files and online appendices are available at https://doi.org/10.1287/mnsc.2022.4507.

Keywords: free trials • targeting • personalization • policy evaluation • field experiment • machine learning • digital marketing • Software as a Service

1. Introduction

Over the last few years, one of the big trends in the software industry has been the migration of software firms from the perpetual licensing business model to the "Software as a Service" (SaaS) model. In the SaaS model, the software is sold as a service; that is, consumers can subscribe to the software based on monthly or annual contracts. Global revenues for the SaaS industry now exceed 200 billion USD (Gartner 2019). This shift in the business model has fundamentally changed the marketing and promotional activities of software firms. In particular, it has allowed firms to leverage a new type of customer acquisition strategy: free trial promotions, where new users get a limited time to try the software for free.

Free trials are now almost universal in the SaaS industry because software is inherently *experience good*, and free trials allow consumers to try the software product without risk. However, we do not have a good understanding of how long these trials should be or the exact mechanism through which they work. In the industry, we observe trial lengths ranging anywhere from one week to three months; for example, Microsoft 365 offers a 30 days free trial, whereas Google's G Suite offers a 14-day free trial. There are pros and cons associated with both long and short trials. A short trial period is less likely to lead to free-riding or demand cannibalization and is associated with lower acquisition costs. Conversely, an extended trial period can enhance consumer learning by giving consumers more time to learn about product features and functionalities. Longer trials can also create stickiness/engagement and increase switching-back costs. That said, if users do not use the product more with a longer trial, they are more likely to conclude that the product is not useful or forget about it. Thus, longer trials lack the deadline or urgency effect (Zhu et al. 2018).

Although the previous arguments make a global case for shorter/longer trials, the exact mechanism at work and the magnitude of its effect can be heterogeneous across consumers. In principle, if there is significant

heterogeneity in consumers' response to the length of free trials, SaaS firms may benefit from assigning each consumer a different trial length depending on her/his demographics and skills. The idea of personalizing the length of free trial promotions is akin to third-degree price discrimination because we effectively offer different prices to different consumers over a fixed period. Indeed, SaaS free trials are particularly well suited to personalization because of a few reasons. First, software services have zero marginal costs, and there are no direct cost implications of offering different trial lengths to different consumers. Second, it is easy to implement a personalized free trial policy at scale for digital services, unlike physical products. Finally, consumers are less likely to react adversely to receiving different trial lengths (unlike prices). However, it is not clear whether personalizing the length of free trials improves customer acquisition and firm revenues, and if yes, what is the best approach to design and evaluate personalized free trials.

In this paper, we are interested in understanding the role of trial length on customer acquisition and profitability for digital experience goods. We focus on the following research questions. First, does the length of a free trial promotion affect customer acquisition, and if so, what is the ideal trial length? Second, what is the mechanism through which trial length affects conversions? Third, is there heterogeneity in users' responsiveness to trial lengths? If yes, how can we personalize the assignment of trial lengths based on users' demographics and skills, and what are the gains from doing so? Furthermore, what types of customers benefit from shorter trials versus longer trials? Finally, how do personalized targeting policies that maximize short-run outcomes (i.e., customer acquisition) perform on long-run metrics such as consumer retention and revenue?

To answer our research questions, we present a three-pronged a three-pronged framework to design and evaluate personalized targeting policies with data from a large-scale free trial experiment conducted by a major SaaS firm. The firm sells a suite of related software products (e.g., Microsoft 365, Google G Suite) and is the leading player in its category, with close to monopolistic market power. At the time of this study, the firm used to give users a 30-day free trial for each of its software products, during which they had unlimited access to the software suite. Then, the firm conducted a large-scale field experiment, where new users who started a free trial for one of the firm's products were randomly assigned to one of 7-, 14-, or 30-day trial length conditions. It also monitored the subscription and retention decisions of the users in the experiment for two years. The firm also collected data on users' pretreatment characteristics (e.g., skill level and job) and posttreatment product usage during the trial period.

First, we quantify the average treatment effect of trial length on subscription. We find that the firm can do significantly better by simply assigning the 7-day trial to all consumers (which is the best uniform policy). This leads to a 5.59% gain in subscriptions over the baseline of 30 days for all policy in the test data. In contrast, the 14 days for all policies does not significantly increase subscriptions. This finding suggests that simply shortening the trial length to 7 days will lead to higher subscriptions. At the time of the experiment, the firm offered a standard 30-day free trial to all its consumers. Therefore, better performance of the much shorter 7-day trial was surprising, especially because the reasons proposed in the analytical literature for the efficacy of free trials mostly support longer trials, for example, switching costs, consumer learning, software complexity, and signaling. (See Section 2 for a detailed discussion of the analytical literature on free trials.) Therefore, we next examine the mechanism through which trial length affects conversion and present some evidence for why a shorter trial works better in this setting and examine the generalizability of these results. To that end, we leverage the usage data during the trial period to understand the mechanism through which trial length affects subscriptions. We show that there are two opposing effects of trial length. On the one hand, as trial length increases, product usage and consumer learning about the software increases. This increase in usage can have a positive effect on subscriptions. On the other hand, as trial length increases, the gap between the last active day and the end of the trial increases, while the average number of active days and usage per day reduces. These factors are associated with lower subscriptions. In our case, the latter effect dominates the former, and shorter trials are better.

Our analysis presents three key findings relevant to the theories on the role of free trials for experience goods. First, we rule out the demand cannibalization or free riding hypothesis advocated by many theoretical papers by showing that users who use the product more during the trial are more likely to subscribe (Cheng and Liu 2012). Second, we provide empirical support for the consumer learning hypothesis, because we show that longer trials lead to more usage, which in turn is associated with higher subscriptions (Dey et al. 2013). Third, we identify a novel mechanism that plays a significant role in the effectiveness of free trials: the negative effect of long stretches of inactivity at the end of the trial on subscription.

Next, we develop a two-step approach to personalized policy design because an unstructured search for the optimal policy is not feasible in our high-dimensional setting. In the first stage, we learn a lasso model of outcomes (subscription) as a function of the users' pretreatment demographic variables and their trial

length. Then in the second stage, we use the individual-level predictions of the outcome to assign the optimal treatment for each user. Then, we use the Inverse Propensity Score (IPS) reward estimator, popular in the counterfactual policy evaluation literature in computer science, for offline policy evaluation (Horvitz and Thompson 1952, Dudík et al. 2011).

Based on this approach, we show that the personalized free trial policy leads to more than 6.8% improvement in subscription compared with the baseline uniform policy of giving a 30-day trial for all. That said, the magnitude of gains from personalization (over the best uniform policy of 7 days for all) are modest (which is in line with the recent findings on personalization of marketing interventions in digital settings; Rafieian and Yoganarasimhan 2021). Furthermore, we find that customers' experience and skill level affect their usage, which affects their subscription patterns. Beginners and inexperienced users show only a small increase in usage with longer trial periods. Furthermore, when given longer trials, they end up with long periods of inactivity at the end of the trial period, which negatively affects their likelihood of subscribing. Thus, it is better to give them short trials. In contrast, long trials are better for experienced users because it allows them to use the software more, and they are not as negatively influenced by periods of inactivity later in the trial period. Overall, our findings suggest that simpler products and experienced users are more likely to benefit from longer trials.

Next, we find that the personalized policy, designed to optimize subscriptions, also performs well on long-term metrics, with a 7.96% increase in customer retention (as measured by subscription length) and 11.61% increase in revenues. We also consider two alternative personalized policies designed to maximize subscription length and revenues and compare their performance with that of the subscription-optimal policy. Interestingly, we find that the subscription-optimal policy always performs the best, even on long-run outcomes. Although this finding is specific to this context, it nevertheless shows that optimizing low-variance intermediate outcomes (i.e., statistical surrogates) can be revenue or loyalty optimal in some settings.

Finally, we consider counterfactual policies based on four other outcome estimators: (1) linear regression, (2) CART, (3) random forests, and (4) XGBoost, and two heterogeneous treatment effect estimators: (1) causal tree, and (2) generalized random forests. We find our lasso-based personalized policy continues to perform the best, followed by the policy based on XGBoost (6.17% improvement). However, policies based on other outcome estimators (e.g., random forests, regressions) perform poorly. Interestingly, policies based on the recently developed heterogeneous treatment effects estimators (causal tree and causal forest) also perform poorly. Causal tree is unable to personalize the policy at all. Causal forest personalizes policy by a small amount, but the gains from doing so are marginal. Although our findings are specific to this context, it nevertheless suggests that naively using these methods to develop personalized targeting policies can lead to suboptimal outcomes. This is particularly important because these methods are gaining traction in the marketing literature and are being used without evaluation using off-policy methods (Fong et al. 2019, Guo et al. 2021).

Our research makes three main contributions to the literature. First, from a substantive perspective, we present the first empirical study that establishes the causal effect of trial length on conversions and provides insight into the mechanisms at play. Second, from a methodological perspective, we present a framework that managers and researchers can use to design and evaluate personalized targeting strategies applicable to a broad range of marketing interventions. Finally, from a managerial perspective, we show that the policies designed to optimize short-run conversions also perform well on long-run outcomes in our setting and may be worth considering in other similar settings. Importantly, managers should recognize that many popular estimators can give rise to poorly designed personalized policies, which are no better than simple uniform policies. Offline policy evaluation is thus a critical step before implementing any policy.

2. Related Literature

Our paper relates to the research that examines the effectiveness of free trials on the purchase of experience goods, especially digital and software products. Analytical papers in this area have proposed a multitude of theories capturing the pros and cons of offering free trials. Mechanisms such as switching costs, network effects, quality signaling, and consumer learning are often proposed as reasons for offering free trials. In contrast, free-riding and demand cannibalization are offered as reasons against offering free trials. See Cheng and Liu (2012), Dey et al. (2013), and Wang and Özkan-Seely (2018) for further details. Despite this rich theory literature, very few empirical papers have examined whether and how free trials work in practice. In an early paper, Scott (1976) uses a small field experiment to examine whether users given a two-week free trial are more likely to purchase a newspaper subscription. Interestingly, this work finds that free trials do not lead to more subscriptions compared with the control condition. Although the number of participants may not have been sufficient to detect small effects and the context was very different from digital SaaS products, it nevertheless raises the question of whether free trials can be an effective marketing strategy. More recently, two empirical papers study free trials using observational data. Foubert and

Gijsbrechts (2016) build a model of consumer learning and show that, although free trials can enhance adoption, ill-timed free trials can also suppress adoption. Using a bayesian learning approach, Sunada (2018) compares the profitability of different free trial configurations. However, neither of these papers examines how trial length affects subscriptions/revenues because they lack variation in the length of the free trials in their data. In contrast, we use data from a large-scale field experiment with exogenous variation in the length of free trials to identify the optimal trial length for each user. In addition, we contribute to this literature by leveraging the individual-level software usage data during the trial period to rule out some of the earlier theories proposed in this context, for example, free riding. To the best of our knowledge, our paper provides the first comprehensive empirical analysis of how trial length affects the purchase of digital experience goods.

Second, our paper relates to the marketing literature on real-time customization and personalization of digital products and promotions using machine learning methods. This literature has used a wide range of methods for the personalization tasks in a variety of contexts: website design using dynamic programming and adaptive experiments (Hauser et al. 2009), display ads using multiarm bandits (Schwartz et al. 2017), ranking of search engine results using feature engineering, and boosted trees (Yoganarasimhan 2020), mobile ads using behavioral and contextual features (Rafieian and Yoganarasimhan 2021), and the sequence of ads shown in mobile apps using batch reinforcement learning and optimal dynamic auctions (Rafieian, 2019a; b). We add to this literature in two ways. First, we document the gains from personalizing the duration of a new type of promotional strategy: the length of time-limited free trials for digital experience goods using a targeting framework based on data from a large-scale field experiment. Second, we show that, although personalization can help, it may not always be the case. Indeed, in our setting, many commonly used methods for personalization often perform worse than a robust uniform policy based on average treatment effects. Although these findings are specific to our context, it nevertheless suggests that managers should be careful in designing and evaluating personalized targeting policies.

Our paper also relates to the theoretical and empirical research on personalized policy design and evaluation in computer science and economics. In an early theoretical paper, Manski (2004) presents a method that finds the optimal treatment for each observation by minimizing a regret function. Recent theoretical papers in this area include Swaminathan and Joachims (2015), Swaminathan et al. (2017), Kitagawa and Tetenov (2018), and Athey and Wager (2020). There is also a small but growing list of marketing papers in this area. Hitsch and Misra (2018) propose a heterogeneous treatment effects estimator based on k Nearest Neighbors, develop targeting policies based on it, and evaluate the performance of their policies using the IPS estimator on a test data. However, their estimator does not work in our setting because it requires all the covariates to be continuous since it based on Euclidean distance. Simester et al. (2020a) examine how managers can evaluate targeting policies efficiently. They compare two types of randomization approaches: (a) randomization by action and (b) randomization by policy. The provide two valuable insights. First, they note that randomization by action is preferable to randomization by policy because it allows us use off policy evaluation to evaluate any policy. Second, they note that when comparing two policies we should recognize that if the policies recommend the same action for some customers then the difference in the performance of the policy for these customers is exactly zero. In another particularly relevant paper, Simester et al. (2020b) investigate how data from field experiments can be used to design targeting policies for new customers or new regimes, and also use the IPS estimator to evaluate the peformance of a series of personalized policies. They present comparisons for a broad range of methods and show that model-based methods in general (and lasso in particular) offers the best performance, though this advantage vanishes if the setting and/or consumers change significantly. Our paper also echoes this finding: the lasso-based personalized policy performs the best in our setting too. Furthermore, we also provide comparisons to personalized policies based on the newly proposed heterogeneous treatment effects estimators (e.g., causal forest) and show that the lasso-based policy continues to perform the best.

Our paper is relevant to the literature on statistical surrogates (Prentice 1989, VanderWeele 2013). In our setting, subscription can be interpreted as an intermediate outcome or surrogate for long-run retention and revenue. Interestingly, we find that personalized policies optimized on the short-term outcome or surrogate do well (or better than) policies optimized directly on the long-term outcomes. We attribute this to the fact that long-term outcomes have higher variance and fewer observations in our setting. In a recent paper, Yang et al. (2022) use surrogates to impute the missing long-term outcomes and then use the imputed long-term outcomes to develop targeting policies. Their results confirm our broader finding that short-term outcomes can be sufficient to derive targeting policies that are optimal from a long-run perspective.

More broadly, our work relates to the large stream of marketing literature that has examined and contrasted the short versus long run effects of promotions (Mela et al. 1997, Pauwels et al. 2002). The main takeaway from this literature is that consumers who are exposed to frequent price promotions become price sensitive and engage in forward buying over time.

Although these early papers focused on consumer packaged goods, Anderson and Simester (2004) conduct a field experiment on price promotions in the context of durable goods sold through catalogs. They find evidence in support of both forward buying and increased deal sensitivity. Our paper adds to this literature by examining the long-run effect of free-trial promotions on long-run subscription and revenue for digital SaaS products. Although free-trial promotions can be viewed as a price discount (i.e., zero price for a fixed period), forward buying is not feasible in our setting, and consumers are exposed to the promotion only during the sign-up period (i.e., no expectation of future free trials). In this case, we find that targeted free-trial promotions that maximize short-run revenue (or subscriptions) also perform well on long-run outcomes (two-year revenue).

3. Setting and Data

In this section, we describe our application setting and data.

3.1. Setting

Our data come from a major SaaS firm that sells a suite of software products. The suite includes a set of related software products (similar to Excel, Word, PowerPoint in Microsoft's MS Office). The firm is the leading player in its category, with close to monopolistic market power. Users can either subscribe to single-product plans that allow them access to one software product or to bundled plans that allow them to use several products at the same time. Bundles are designed to target specific segments of consumers and consist of a set of complementary products. The prices of the plans vary significantly and depend on the bundle, the type of subscription (regular or educational), and the length of commitment (monthly or annual). Standard subscriptions run from \$30 to \$140 per month depending on the products in the bundle and come with a monthly renewal option. (To preserve the firm's anonymity, we have multiplied all the dollar values in the paper by a constant number.) If the user is willing to commit to an annual subscription, they receive more than a 30% discount in price. However, users in annual contracts must pay a sizable penalty to unsubscribe before the end of their commitment. The firm also offers educational licenses at a discounted rate to students and educational institutions, and these constitute 20.8% of the subscriptions in our data.

3.2. Field Experiment

At the time of this study, the firm used to give users a 30-day free trial for each of its software products, during which they had unlimited access to the product.[1] To access the product after the trial period, users need a subscription to a plan or bundle that includes that product.

To evaluate the effectiveness of different trial lengths, the firm conducted a large-scale field experiment that ran from December 1, 2015, to January 6, 2016, and spanned six major geographic markets: Australia and New Zealand, France, Germany, Japan, the United Kingdom, and the United States. During the experiment period, users who started a free trial for any of the firm's four most popular products were randomly assigned to one of 7-, 14-, or 30-day free trial length buckets. These three trial lengths were chosen because they are the most commonly used ones in the industry and represent a vast majority of the SaaS free trials. Treatment assignment was at user level, that is, once a user was assigned to a trial length, her/his trial length for the other three popular products was also set at the same length. The length of the free trial for other products during this period remained unchanged at 30 days. The summary statistics for the treatment assignment and subscriptions are shown in Table 1.

The experiment was carefully designed and implemented to rule out the possibility of self-selection into treatments, a common problem in field experiments. In our setting, if users can see which treatment (or free trial length) they are assigned to prior to starting their trial, then users who find their treatment undesirable may choose to not start the trial. In that case, the observed sample of users in each treatment condition would no longer be random, and this in turn would bias the estimated treatment effects. Moreover, because the experimenter cannot obtain data on those who choose to not to start their free trials, there is no way to address this problem econometrically. To avoid these types of self-selection problems, the firm designed the experiment so that users were informed of their trial length only after starting their trial. To try a software product, users had to take the following steps: (1) sign up with the firm by creating an ID, (2) download an app manager that manages the download and installation of all the firm's products, and (3) click on an embedded *start trial* button to start the trial for a given product. Only at this point in time are they shown the length of their free trial as the time left before their trial expires (e.g., "Your free trial expires in 7 days"). Although users can simply quit or choose to not use the product at this point, their identities and actions are nevertheless captured in our data and incorporated in our analysis.

Finally, it is important to note that treatment assignment was unconfounded with other marketing mix variables. In this context, it is useful to discuss prices because they can vary across products and users. The price that a user gets for a product/bundle depends only on two user-level observables: the geographic

Table 1. Summary Statistics of Treatment Assignment and Subscription Rates

	7-day trial	14-day trial	30-day trial	Total
Number of observations (N)	51,040	51,017	235,667	337,724
Percent of total observations	15.11	15.11	69.78	100
Number of subscriptions	7,835	7,635	34,564	50,034
Percent of total subscriptions	15.66	15.26	69.08	100
Subscription rate within group (in %)	15.36	14.96	14.67	14.81

location of the user and her/his job (students get a discount). Both variables are observed in the data, and treatment assignment is orthogonal to these variables. Thus, price is not confounded with treatment.[2]

In sum, the design satisfies the two main conditions necessary for the experiment to be deemed clean: (1) unconfoundedness and (2) compliance (Mutz et al. 2019).

3.3. Data

We have data on 337,724 users who were part of the experiment. For each user i, we observe the following information: (1) treatment assignment (W_i), (2) pretreatment demographic data (X_i), and (3) posttreatment behavioral data (Z_i). The treatment variable denotes the trial length that the user was assigned to: 7, 14, or 30 days. The variables under the latter two categories are described in detail here.

3.3.1. Pretreatment Demographic Data.

1. Geographic region: The geographic region/country that the user belongs to (one of the six described in Section 3.1). It is automatically inferred from the user's IP address.

2. Operating system: The OS installed on the user's computer. It can take eight possible values, for example, Windows 7, Mac OS Yosemite. It is inferred by the firm based on the compatibility of the products downloaded with the user's OS.

3. Sign-up channel: The channel through which users came to sign-up for the free trial. In total, there are 42 possible sign-up channels, for example, from the legacy version of the software, from the firm's website, through third-parties, and so on.

4. Skill: A self-reported measure of the user's fluency in using the firm's software suite. This can take four possible values: beginner, intermediate, experienced, and mixed.

5. Job: The user's job-title (self-reported). The firm gives users 13 job-titles to pick from, for example, student, business professional, hobbyist.

6. Business segment: The self-reported business segment that the user belongs to. Users can choose from six options here, for example, educational institution, individual, enterprise.

The last three variables are self-reported although not open-ended; that is, users are required to pick one option from a list provided by the firm. However, users may choose not to report these values, in which case, the missing values are recorded as "unknown." We treat this as an additional category for each of these three variables in our analysis.[3] The six pretreatment variables and their summary statistics are shown in Table 2.

3.3.2. Posttreatment Behavioral Data.

For all the users in our data, we observe their subscription and renewal decisions for approximately 24 months (from December 2015 until November 2017). We have data on the following:

1. Subscription information: We have data on whether a user subscribes or not, and the date and type of subscription (product or bundle of products) if she does subscribe.

2. Subscription length: Number of months that the user is a subscriber of one or more products/bundles during the 24-month observation period. If a user does not subscribe to any of the firm's products during the observation period, then this number is zero by default.[4]

Table 2. Summary Statistics for the Pretreatment Categorical Variables

		Share of top subcategories			
Variable	Number of subcategories	First	Second	Third	Fourth
Geographic region	6	55.02%	13.66%	9.12%	8.83%
Operating system	8	28.97%	21.4%	14.04%	13.98%
Signup channel	42	81.56%	8.14%	3.47%	0.81%
Job	14	28.20%	21.90%	20.34%	8.46%
Skill	5	69.05%	12.75%	10.77%	7.38%
Business segment	7	35.41%	32.74%	18.40%	7.81%

3. Revenue: The total revenue (in scaled dollars) generated by the user over the two-year observation period. This is a function of the user's subscription date, the products and/or bundles that she subscribes to, the price that she pays for her subscription, and subscription length.

The summary statistics of these outcome variables are presented in Table 3. Both subscription length and revenue are shown for (a) all users and (b) the subset of users who subscribed. There are a couple of points to note here. First, we do not have access to the subscription length and revenue data for team subscriptions and government subscriptions (which constitute a total of 3,501 subscriptions). Hence, the number of observations used to calculate the summary statistics for subscription length and revenue for subscribers is lower. Second, the minimum subscription length observed in the data for subscribers is zero because we have a few users (58 users) who immediately unsubscribed after subscribing (within one month), in which case the firm returns their money and records their subscription length and revenue as zero. Based on Table 3, we see that approximately 14.8% users who start a free trial subscribe, and the average subscription length of subscribers is about 16 months (which is a little over a year).

We also observe the following product download and usage data for the duration of a user's trial period.

1. Products downloaded: The date and time-stamp of each product downloaded by the user.

2. Indicator for software use: An indicator for whether the user used the software at all.

3. Number of active days: Total number of days in which the user used the software during the trial period. For example, if a user with a 7-day trial uses the software on the first and third day, this variable is two.

4. Usage during trial: Each product in the software suite has thousands of functionalities. Functionalities can be thought of as microtasks and are defined at the click and key stroke level; for example, save a file, click undo, and create a table. The firm captures all this information, and we have data on the total count of the functionalities used by the user during her trial period.

5. Dormancy length: The number of days between the last active day and the last day of trial, as shown in Figure 1. For example, if a user with a 30-day trial last used the software on day 20, then her dormancy length is 10.

We present the summary statistics of these usage variables in Table 4. The usage data are also missing (at random) for a subset of users, and we report the summary statistics for nonmissing observations. As we can see, most users download only one software product; only 13.6% of people download more than one product. Furthermore, 83% of users try the software at least once. However, the number of active days is relatively small; the average user uses the software for only three days during the trial period. Next, we see that an average user uses 1,733 functionalities during her trial. However, notice that this variable is very skewed, with the variance being much higher than the mean. Therefore, we use the natural log of this variable in all our analyses going forward. Finally, we see that the average dormancy length is close to 17 days, which means that many users stop using the software much before the end of trial period.

Finally, we refer interested readers to Tables A1 and A2 in Online Appendix A for the summary statistics of outcome and usage variables by trial length, respectively.

3.3.3. Training and Test Data.

To design and test counterfactual free trial policies, we partition the data into two independent samples.

- **Training Data:** This is the data that is used for both learning the model parameters and model selection (or hyper-parameter optimization through cross-validation).
- **Test Data:** This is a hold-out data on which we can evaluate the performance of the policies designed based on the models built on training data.

We use 70% of the data for training (and validation) and 30% for test. See Table A3 in Online Appendix A for a detailed breakdown of how the data are split across the two data sets. Although the joint distributions of the variables in the two samples should be the same theoretically, there will be some minor differences between the two data sets because of the randomness in splitting in a finite sample. It is important to keep this in mind when comparing results *across* the two data sets.

4. Main Effect of Trial Length on Subscription

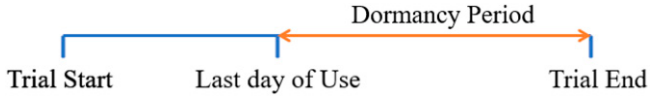
We now document the main effect of trial length on subscription and present some evidence for the mechanism

Table 3. Summary Statistics of Subscription, Subscription Length, and Revenue Outcomes

Variable	Mean	Standard deviation	Minimum	25%	50%	75%	Maximum	Number of observations
Subscription	0.148	0.355	0	0	0	0	1	337,724
Subscription length (all)	2.23	6.37	0	0	0	0	108	334,223
Subscription length (subscribers)	16.02	8.43	0	10	17	22	108	46,533
Revenue (all)	79.13	285	0	0	0	0	20,208	334,223
Revenue (subscribers)	568	552	0	242	420	666	20,208	46,533

Note. All the revenue numbers are scaled by a constant factor to preserve the firm's anonymity.

Figure 1. (Color online) Dormancy Length: Number of Days Between the Last Active Day and the End of the Trial Period



behind this effect. For expositional simplicity, we focus on subscription here and present a detailed analysis long-run outcomes such as revenue and retention in Section 7.

4.1. Average Treatment Effect

In a fully randomized experiment (such as ours), the average effect of a treatment can be estimated by simply comparing the average of the outcome of interest across treatments. We set the 30-day condition as the control and estimate the average effects of the 14- and 7-day trials on subscriptions for training and test data. The results from this analysis are shown in Table 5.

The 7-day trial increases the subscription rate by 4.34% over the baseline of the 30-day condition in the training data and by 5.59% in the test data. However, in both data sets, the effect of the 14-day trial is not significantly different from that of the 30-day trial. These results suggest that a uniform targeting policy that gives the 7-day treatment to all users can significantly increase subscriptions.[5] We also see that the average treatment effect is fairly small compared with the outcome, which is either zero or one. This is understandable because the effect of trial length is likely to be small compared with other factors that affect customer acquisition. Finally, the gains and subscription rates in the training and test data are slightly different. As discussed earlier, this is because of the randomness in the splitting procedure.

Next, to ensure that these results are not driven by any problems with randomization, we conduct a series of randomization checks. We present the details of these tests in Online Appendix B.2 and discuss them briefly here. First, we conduct a joint test of orthogonality of pretreatment variables and treatment assignment (McKenzie 2017). This is done by regressing the treatment variable on the entire set of pretreatment variables (with dummies for each subcategory shown in Table 2). We find that the pretreatment variables have no predictive power when it comes to predicting treatment, which suggests that randomization was done correctly. This approach to checking for potential issues with randomization is preferable to the old practice of showing tables of means for pretreatment variables across treatment arms and running a battery of t tests for a variety of reasons (see Bruhn and McKenzie (2009) and Mutz et al. (2019) for detailed discussions).[6] Next, we regress the outcome variable (subscription outcome) on the treatment variable and all the pretreatment variables. We find that the treatment effects are very similar to those in Table 5, which again suggests that there are no issues with randomization.[7]

4.2. Mechanism

At the time of the experiment, the firm offered a standard 30-day free trial to all its consumers. The better performance of the much shorter 7-day trial was both surprising and inexplicable for many reasons. First, the firm sells a complicated suite of software with multiple products and functionalities. Therefore, we would have expected that giving consumers more time to familiarize themselves with it and learn the software would produce better outcomes. Second, the reasons proposed in the theory literature for the efficacy of free trials largely support longer free trials, for example, switching costs, consumer learning, software complexity, and signaling. Thus, it is not obvious why a shorter trial works better. Therefore, we now examine how trial length affects conversion and present some evidence for why a shorter trial works better in this setting. In the process, we also discuss the generalizability of our findings and the mechanisms proposed.

Intuitively, trial length can affect how consumers download, use, and interact with the software; differences in these usage variables can lead to different subscription outcomes. Therefore, we first examine whether and how trial length affects usage. We regress each of

Table 4. Summary Statistics for Usage Features

Variable	Mean	Standard deviation	Minimum	25%	50%	75%	Maximum	N
Total downloaded packages	1.17	0.41	1.0	1.00	1.00	1.00	4.00	337,724
Indicator for software use	0.83	0.37	0.0	1.00	1.00	1.00	1.00	303,514
Number of active days	3.03	3.94	0.0	1.00	2.00	4.00	30.00	303,514
Usage during trial	1,733	7,220	0	47	257	1,086	488,666	303,514
Log usage during trial	5.09	2.74	0.0	3.87	5.55	6.99	13.10	303,514
Dormancy length	16.87	11.23	0.0	6.00	15.00	29.00	30.00	303,514

Table 5. Average Effect of the 7- and 14-Day Treatments on Subscription Compared with the Control Condition of 30-Day Free Trial

Data	Treatment	Subscription rate	Subscription rate difference	t statistics	Percentage gain over baseline
Training data	7 days	0.1532	0.0064	3.08	4.34
	14 days	0.1490	0.0021	1.03	1.45
	30 days	0.1468	—	—	—
Test data	7 days	0.1544	0.0082	2.58	5.59
	14 days	0.1511	0.0048	1.51	3.28
	30 days	0.1463	—	—	—

Note. Baseline subscription rate (for 30-day case): 14.68 in training data and 14.63 in test data.

the usage variables shown in Table 4 on trial length and present the results in Table 6. Because trial length is randomly assigned, we can interpret these results causally. First, we find that longer free trials lead to more product downloads and more usage. Further investigation suggests that this increase in downloads mainly comes from the higher downloads of products 1 and 3, which are complements (see Figure A1 in Online Appendix C). This suggests that giving longer trial lengths to users increases their probability of exploring other complementary products. Next, we see that a larger fraction of people try the software at least once with a longer trial, and the number of active days and log usage also increases with trial length. However, the rate of increase in the number of active days and usage is sublinear compared with the increase in trial length. For instance, going from 7 to 14 days increases the number of active days by 0.625, which is much smaller than 7 days (the increase in trial length). Thus, when we normalize the number of active days by trial length, the average number of days during which a user is active during her trial reduces as trial length increases. The same pattern holds for log usage; although total usage increases as trial length increases, average daily usage falls. Finally, we find that the dormancy period increases as trial length increases. Although the average dormancy length is 4.6 days for the 7-day trial, it is more than 21 days for the 30-day trial.

Next, we examine whether and how usage is associated with subscription. The left panel of Figure 2 shows the probability of subscription as a function of the total number of active days for each trial length. As we can see, users who are active for more days are also more likely to subscribe, and this pattern is true for all three trial lengths. However, given the same level of activity, shorter trial lengths are associated with higher conversion. For example, users who were active for five days are more likely to subscribe when they are in the 7-day condition compared with the 14- or 30-day condition. Next, in the right panel of Figure 2, we show the probability of subscription as a function of the last active day for all three trial lengths. We see that users whose last active day is earlier in the trial period are less likely to subscribe. Furthermore, for the same last active day, users with shorter trials are more likely to subscribe. Recall that dormancy length is defined as trial length minus the last active day. Therefore, this suggests that users who have not used the product for long periods at the end of the trial period are less likely to subscribe.

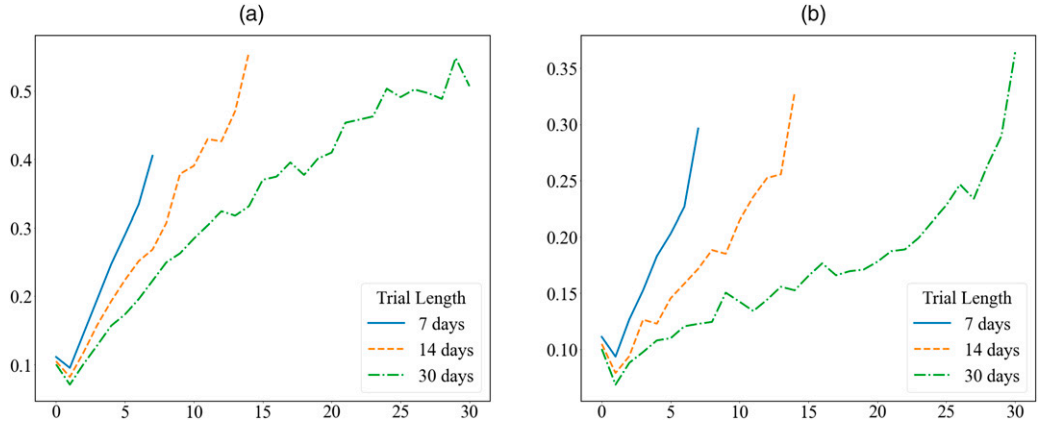
We now check if the preliminary patterns shown in Figure 2 hold after we control for other usage and user-specific observables. In Table 7, we present the results from a regression with the user's subscription decision as the outcome variable and her trial length and usage variables as explanatory variables. We find that after controlling for everything else, users who have more active days and use the product more are more likely to subscribe. Furthermore, users who have longer dormancy periods are less likely to subscribe. This is understandable because a user who has not used the software for a long time by the end of her trial period is likely to forget about it and/or conclude that the product is not useful (Zhu et al. 2018).

Table 6. Regression of Usage Features on Trial Length

Outcome variable	Intercept	14-day trial	30-day trial	R^2	N
Total downloaded packages	1.137 (0.002)	0.01 (0.002)	0.017 (0.002)	0.000	337,724
Indicator for software use	0.828 (0.002)	0.004 (0.002)	0.009 (0.002)	0.000	303,514
Number of active days	1.747 (0.018)	0.625 (0.026)	1.711 (0.02)	0.028	303,514
Number of active days/trial length	0.25 (0.001)	−0.08 (0.001)	−0.134 (0.001)	0.078	303,514
Log usage during trial	4.77 (0.013)	0.196 (0.018)	0.411 (0.014)	0.003	303,514
Log average daily usage during trial	3.197 (0.009)	−0.357 (0.012)	−0.737 (0.009)	0.022	303,514
Dormancy length	4.631 (0.043)	5.135 (0.06)	16.432 (0.047)	0.337	303,514

Note. Standard errors in parentheses.

Figure 2. (Color online) Relationship Between Usage and Subscription Rate for Different Trial Lengths



Notes. (a) The subscription rate based on the last day of trial use for different trial lengths. (b) The subscription rate based on the number of active days for different trial lengths.

Together, the previous findings suggest that two opposing effects of trial length on usage and subscription. We depict these effects in Figure 3. On the one hand, as trial length increases, product usage and consumer learning about the software increases. This increase in usage can have a positive effect on subscriptions. On the other hand, as trial length increases, the gap between the last active day and the end of the trial increases, whereas the average number of active days and usage per day reduces. These factors are associated with lower subscriptions. In our case, it seems that the latter effect dominates the former, and hence shorter trials are better.[8]

Our analysis presents three key findings relevant to the broader theories on the role of free trials for experience goods. First, we rule out the well-known demand cannibalization hypothesis advocated by many theoretical papers (Cheng and Liu 2012, Dey et al. 2013). These papers argue that, with longer trials, free-riders can use the product extensively during the trial, get their project/job done, and avoid subscribing. However, the results in Figure 2 and Table 7 rule out the free-riding hypothesis because users who use the product heavily during the trial are also more likely to subscribe. However, this evidence is for the full population of users. Second, we provide empirical support for the consumer learning hypothesis proposed in analytical papers (Dey et al. 2013) because we find that longer trials lead to more usage, which in turn is associated with higher subscription. Third, we identify a novel mechanism that plays a significant role in the effectiveness of free trials: the negative effect of long dormancy periods on subscription. We provide more evidence in support of these ideas in Section 6, where we discuss the heterogeneous response of different types of users.

4.3. Heterogeneity in Users' Responsiveness

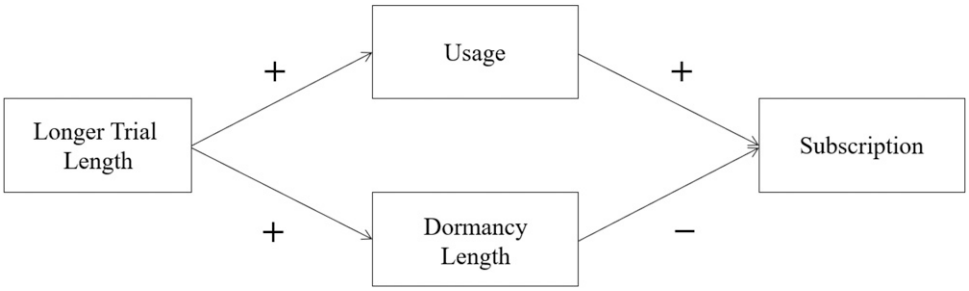
Thus far, we have shown that the 7-day trial is the best average treatment and provided some intuition for why. However, the effect of trial length could be heterogeneous across users and the mechanisms discussed earlier could be differentially important for different types of users. We now examine whether this is indeed the case.

In the top left panel of Figure 4, we partition the data into six subgroups based on the user's geographic region and present the average subscription rates for the three trial lengths for each region. The results suggest that there is some heterogeneity in response rates by region. For example, the 14-day trial is more effective in Germany, whereas the 7-day trial is more effective in the United States. Next, we perform a similar exercise on skill-level and job (see the top right and bottom panels in Figure 4). Again, we

Table 7. Regression of Subscription on Usage Features and Trial Length, with All Pretreatment Variables Included as Controls (Not Shown in Table 6)

	Coefficent	Standard error	z	$P>\|z\|$	[0.025	0.975]
Indicator for using the software	−0.5145	0.036	−14.252	0.000	−0.585	−0.444
Total downloaded packages	0.5632	0.013	43.789	0.000	0.538	0.588
Number of active days	0.0440	0.002	19.241	0.000	0.040	0.049
Log usage during trial	0.0620	0.005	11.267	0.000	0.051	0.073
Dormancy length	−0.0297	0.001	−30.141	0.000	−0.032	−0.028

Figure 3. Effect of Trial Length on Usage and Dormancy Length and Subsequently Subscription



find that users' responsiveness to the treatment is a function of their skill level and job. For instance, the 7-day trial is significantly better for beginners, whereas the 14-day trial is more effective for mixed-skill users.

These results suggest that users' responsiveness to trial lengths is heterogeneous on many pretreatment variables. If the firm can successfully exploit the different sources of heterogeneity and personalize its free trial assignment at the individual level, then it may be able to further improve subscriptions.

5. Counterfactual Analysis: Personalized Policy Design and Evaluation

The previous preliminary evidence suggests that the firm can benefit from personalizing free trial assignment. In Section 5.1, we describe the procedure we use to design the personalized policy. Next, in Section 5.2, we present the gains from the personalized policy in our setting. Next, in Section 5.3, we compare the performance of our approach to other personalized policies.

5.1. Optimal Policy Design

Let $i \in \{1, \ldots, N\}$ denote the set of independent and identically distributed users, where each user is characterized by a pretreatment covariate vector $X_i \in X$ of dimension D. Let $W_i \in \mathcal{W}$ denote the treatment or intervention that i receives. $\mathcal{W} = \{0, \ldots, W-1\}$ refers to the set of treatments, and the total number of treatments is W. Finally, let $Y(X_i, W_i)$ denote the outcome for a user i with pretreatment variables X_i when she is allocated treatment W_i.

A personalized treatment assignment policy, π, is defined as a mapping between users and treatments such that each user is allocated one treatment, $\pi : X \rightarrow \mathcal{W}$. The firm's goal is to choose a policy π such that it maximizes the expectation of outcomes, $\frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} Y(X_i, W_i^{\pi})\right]$. Thus, for policy π and outcome of interest Y, we can write our reward function as $R(\pi, Y) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y(X_i, \pi(X_i))\right]$. Thus, given a reward function $R(\pi, Y)$, the optimal personalized policy is given by

$$\pi^* = \arg\max_{\pi \in \Pi} [R(\pi, Y)], \tag{1}$$

where Π is the set of all possible policies.

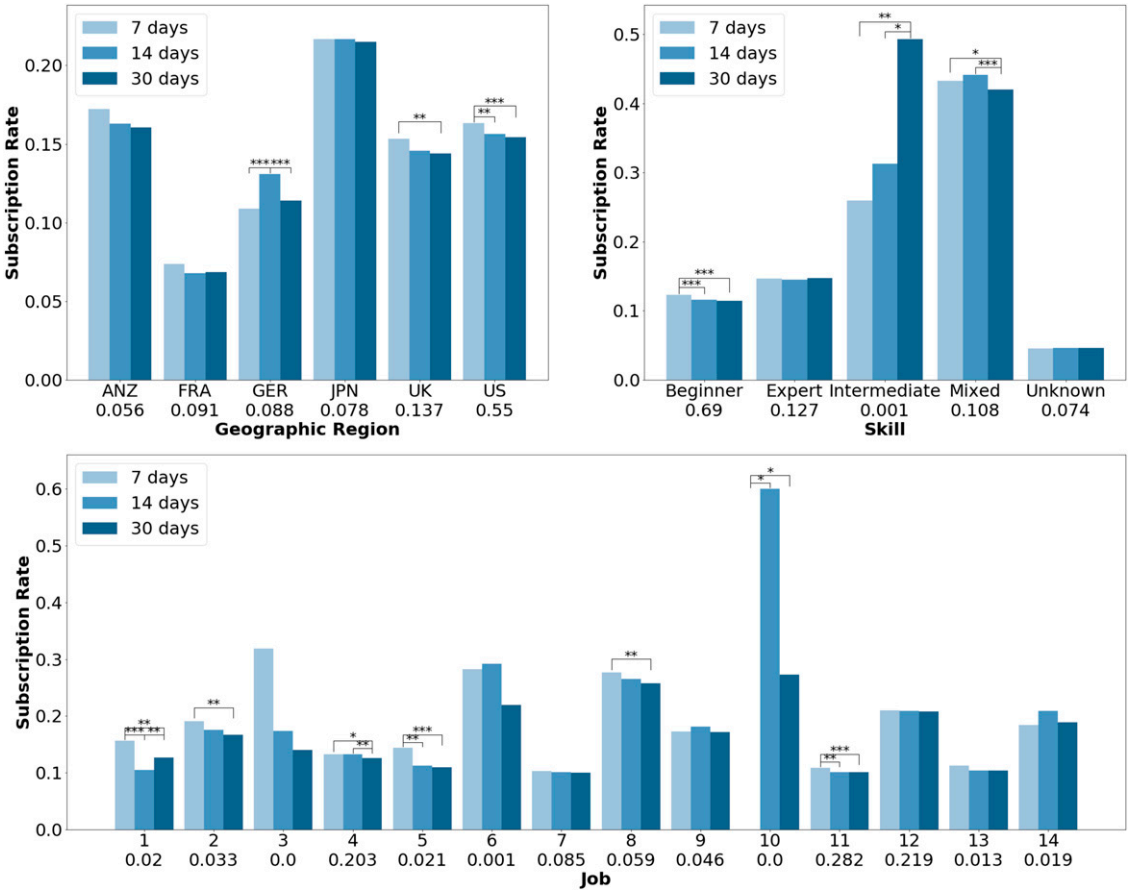
The problem of finding the optimal personalized policy is equivalent to one of finding the policy π^* that maximizes the reward function $R(\pi, Y)$. As discussed in Section 1, this is a nontrivial problem because the cardinality of the policy space can be quite large.[9] Therefore, a direct search over the policy space to find the optimal policy is infeasible. Therefore, we adopt a two-step approach to find the optimal policy π^* that avoids this problem. To do so, we make the three standard assumptions on (1) unconfoundedness, (2) Stable Unit Treatment Value Assumption, and (3) positivity. Given that our data come from a fully randomized experiment, assumptions (1) and (2) are automatically satisfied. Furthermore, assumption 2 is satisfied because we do not expect any network effects in our setting (because the experiment was run on unconnected users distributed all over the world).

With these assumptions in place, we can design the optimal personalized policy if we either have estimates of the outcome of interest or pairwise treatment effects. In the main analysis, we design our personalized policy based using outcome estimates based on lasso (Tibshirani 1996, Friedman et al. 2010). That is, we model the subscription outcome using as $f(x, w) = \mathbb{E}[Y \mid X_i = x, W_i = w]$, where $f(\cdot)$ is a lasso model. Lasso estimates a linear regression that minimizes the MSE with an additional term to penalize model complexity as shown here:

$$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = \arg\min \sum_{i=1}^{n} (Y_i - X_i\beta_1 - W_i\beta_2 - X_iW_i\beta_3)^2 + \lambda(\|\beta_1\|_1 + \|\beta_2\|_1 + \|\beta_3\|_1), \tag{2}$$

where $\|\beta_i\|_1$ is the L1 norm of the vector β_i and is equal to the sum of the absolute value of the elements of vector β_i. Intuitively, if there are multiple weak (and correlated) predictors, lasso will pick a subset of them

Figure 4. (Color online) Heterogeneity in Consumers' Response to the Three Trial Lengths Within Three Categories: Geographic Region, Skill, and Job



Notes. The six geographic regions shown are as follows: Australia and New Zealand, France, Germany, Japan, and United States (in that order). Under each subcategory, the fraction of users in that subcategory are shown. We do not include subcategory names for Job to preserve the firm's anonymity (*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$).

and force the coefficients of others to zero, thereby estimating a simpler model. Model selection in lasso is data-driven; that is, λ is a hyper-parameter that is learned from the data (and not assumed). Please see Online Appendix D for details of our lasso estimation.

Next, using estimate of the expected outcome, $\hat{y}(x = X_i, w)$, from the lasso model, we obtain the optimal personalized policy based on for observation i as

$$\pi_{lasso}(X_i) = w^*, \quad \text{where} \quad w^* = \arg\max_{w \in \mathcal{W}} \hat{y}(x = X_i, w). \tag{3}$$

Our personalized free trial policy, π_{lasso} partitions the population into three segments: 7-, 14-, and 30-days optimal segments, which constitute 68.9%, 23.2%, and 7.9% of the population, respectively.

5.2. Empirical Policy Evaluation: Gains from Personalization

We now empirically evaluate and quantify the gains from the personalized free trial policy π_{lasso} over nonpersonalized policies. To do so, we first define three uniform (one length for all) policies:

- π_{30}: This policy prescribes the 30-day treatment for all users. It was used by the firm at the time of the experiment, and we therefore use it as the baseline policy in all our comparisons.
- π_{14}: This policy prescribes the 14-day treatment for all users.
- π_7: This policy prescribes the 7-day treatment for all users. Because we found that 7 days is the best average treatment in Section 4.1, this is the best uniform policy.

We evaluate the expected reward from the policies (both personalized and uniform) using the IPS estimator that has been extensively used in the off-policy evaluation literature (Horvitz and Thompson 1952, Dudík et al. 2011) and has recently been applied in the marketing too (Hitsch and Misra 2018, Simester et al. 2020b, Rafieian and Yoganarasimhan 2021). For any given policy π, this estimator takes all the observations where the user received the policy-prescribed

treatment and scales them up by their propensity of receiving the treatment assigned to them. This scaling gives us a pseudo-population that received the policy-prescribed treatment. Thus, the average of the outcome for this pseudo-population gives us an unbiased estimate of the reward for the full population, if we were to implement the proposed policy in the field. Formally,

$$\hat{R}_{IPS}(\pi, Y) = \frac{1}{N}\sum_{i=1}^{N}\frac{1[W_i = \pi(X_i)]Y_i}{\hat{e}_{\pi(X_i)}(W_i)}, \tag{4}$$

where $\hat{e}_{\pi(X_i)}(W_i)$ is the probability that a user whom the policy prescribes treatment $\pi(X_i)$ is given W_i.[10]

We present the expected rewards (or subscription rates) from all the three uniform policies and π_{lasso} in the top panel of Table 8. The key finding is that personalization based on pretreatment demographic variables leads to more than 6.8% improvement in subscription compared with the baseline uniform policy of giving a 30-day trial for all. Furthermore, we see that the personalized policy also does better than the best uniform policy of 7 days for all. To examine whether this difference is significant, we conduct a paired t test based on bootstrapping as follows. We repeatedly (20 rounds) split the entire data into training and test (in the same proportion used in the main analysis, that is, 0.7/0.3). Then, in each round, we train a lasso model on the training set using a five-fold cross-validation and calculate the IPS rewards (based on Equation (4)) for both π_7 and π_{lasso} on the test data. Finally, we run a two-sided paired t test to compare lasso's performance with the uniform all 7-day policy. The t statistic and p value for the two-sided test are 3.123 and 0.0056, respectively, which confirms that the personalized policy π_{lasso} is better than the best uniform policy π_7.

That said, notice the magnitude of gains from personalization (over the best uniform policy) is modest. This is understandable because the personalized policy assigns about 70% of users to the 7-day treatment, and the gains from personalization only accrue from the remaining 30% of users who are allocated the 14- or 30-day treatments. As Simester et al. (2020a) point out, this is because the difference in the the performance of the two policies for users assigned to the 7-day trial is exactly zero. Furthermore, our treatment effect is small compared with the outcome: a common occurrence for marketing interventions such as advertising or promotions (Lewis and Rao 2015). These findings are consistent with the recent literature on personalization (Yoganarasimhan 2020, Rafieian and Yoganarasimhan 2021), which demonstrate positive but moderate gains from personalization digital interventions.

5.3. Comparisons and Robustness Checks

Thus far, we used lasso as the outcome estimator to design our personalized policy. We now examine whether counterfactual personalized policies based on other outcome and heterogeneous treatment effects estimators perform better. Specifically, we consider policies based on four additional outcome estimators: (1) linear regression, (2) CART, (3) random forests, and (4) XGBoost, and two conditional average treatment effects (CATE) estimators: (1) causal tree and (2) generalized random forests. The technical details of these models and their tuning details are shown in Online Appendices E and F.

First, we find that each of these policies behaves quite differently when it comes to treatment assignment (see Table A8 in Online Appendix G for details). Interestingly, we find that policies based on CART and causal tree do not personalize treatment assignment and end up giving the 7-day treatment to all users, that is, $\pi_7 \equiv \pi_{cart} \equiv \pi_{c\_tree}$. We also find that in $\pi_{xgboost}$ is quite similar to π_{lasso}. Both prescribe the 7-day trial to $\approx$ 70% of users, the 14-day trial to $\approx$ 20% of users, and the 30-day trial to $\approx$ 10% of users. In contrast, π_{reg} and π_{r_forest} prescribe the 7-day treatment to the least number of users, whereas π_{c_forest} prescribes the 7-day treatment to 91% of users (and the 30-day treatment to no one).

Next, in the bottom panel of Table 8, we present the performance of these policies. We find that π_{lasso} continues to be the best, and the second-best policy is $\pi_{xgboost}$. There are two main takeaways here. First, poorly designed personalized policies (e.g., those based on regression and random forest) can actually do worse than the best uniform policy on the test data. Second, we do not find much correlation between an outcome estimator's predictive ability and its efficacy in policy design. For instance, random forest has a lower mean squared error on the test data compared with lasso, but π_{r_forest} is much worse than π_{lasso} (see Table A9 in Online Appendix G). This is likely because the objective function in outcome estimation methods is predictive ability, which is different from policy design or performance. In sum, our findings suggest that managers should be careful in both designing and evaluating personalized policies. It is critical to (1) not conflate a model's predictive ability with its ability to form policy and (2) evaluate the performance of each policy on an independent test data with appropriate policy evaluation metrics.

Next, we find that the recently proposed heterogeneous treatment effects estimators, causal tree and causal forest, perform poorly when it comes to personalized policy design. Our results suggest that managers may be better off adopting the best uniform policy instead of investing resources in personalizing policies based on these methods. This is an important

Table 8. Gains in Subscription from Implementing Different Counterfactual Free-Trial Policies

		Estimated subscription (%)		Increase in subscription (%)	
Policy category	Policy	Training set	Test set	Training set	Test set
Personalized based on lasso	π_{lasso}	15.85	15.62	7.97	6.81
Uniform	π_7	15.32	15.44	4.34	5.59
	π_{14}	14.90	15.11	1.45	3.28
	π_{30} (Baseline)	14.68	14.63	—	—
Alternative personalized policies	π_{reg}	15.89	15.33	8.21	4.83
	π_{cart}	15.32	15.44	4.34	5.59
	π_{r_forest}	17.42	14.82	18.67	1.32
	$\pi_{xgboost}$	16.00	15.53	8.98	6.17
	π_{c_forest}	15.58	15.46	6.09	5.71
	π_{c_tree}	15.32	15.44	4.34	5.59

Note. The results for policies π_{cart}, π_{c_tree}, and π_7 are the same since they prescribe the 7-day treatment to all users.

finding because these methods are gaining traction in the marketing literature and researchers are starting to use them (Fong et al. 2019, Guo et al. 2021). Our findings suggest that relying on heterogeneous treatment effects estimators can be suboptimal.

Finally, we examine where there is any relationship between the estimates of treatment effects based on a specific method and the performance of the policy based on it. Figure 5 shows the cumulative density function (cdf) of $\tau_{7,30}$ for all the methods used for policy design. The estimated distributions of $\tau_{14,20}$, and $\tau_{7,14}$ are shown in Figure A2 in Online Appendix G, and their interpretations are largely similar to that presented here for $\tau_{7,30}$.[11] The first pattern that stands out is that treatment effects estimates based on CART, causal tree, and casual forest show very little heterogeneity (see the three vertical lines to the right of zero). This explains why policies based on these methods perform poorly; they are unable to personalize the policy sufficiently to optimize users' response at the individual level. In contrast, the treatment effect estimates based on linear regression and random forest show the maximum amount of heterogeneity (see the two rightmost curves). This pattern, in combination with the poor performance of π_{reg} and π_{r_forest} on test data (and their extremely good performance on training data) hints at overfitting problems. That is, these models seem to infer much more heterogeneity than is true in the data. Interestingly, we see that the CDFs of treatment effects based on lasso and XGBoost lie somewhere in between the above two groups. They show sufficient heterogeneity but not too much. Hence, policies based on these methods can personalize the treatment sufficiently without overfitting. Recall that the dispersion in treatment assignment for these two policies is higher than that in π_{cart}, π_{c_tree}, and π_{c_forest}, but lower than that in π_{reg} and π_{r_forest}. Thus, the ideal estimators for policy design are those that are able to capture sufficient heterogeneity to personalize effectively without overfitting (i.e., capture spurious heterogeneity).

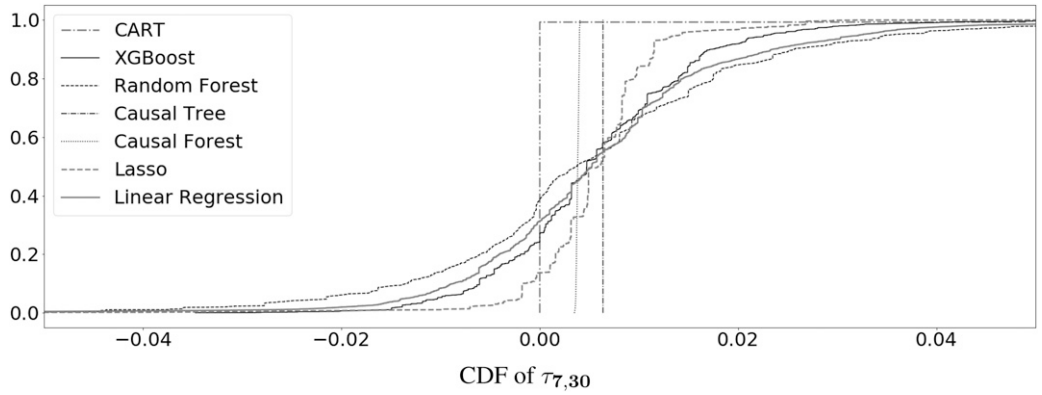
6. Segmentation Analysis and Additional Evidence for Mechanism

Thus far, we focused on the question of "Who (should get a treatment)." We now examine the question of "Why (should s/he get a specific treatment)." Understanding why some users respond well to longer trials, whereas others respond better to shorter trials can give us insight into consumers' preferences and decision-making process. These insights are valuable for two reasons. First, from the firm's perspective, they can be leveraged to improve other marketing interventions such as advertising and pricing. Second, from a research perspective, this gives us a better understanding of the sources of heterogeneity in the effectiveness of trial length on conversion and mechanisms at play, which can be generalized to other settings.

We now correlate a user's optimal treatment with her pretreatment demographic and posttreatment behavioral variables. In the process, we present additional evidence for the mechanism through which trial length affects conversion, as discussed in Section 4.2. We conduct three sets of analyses to understand the mechanism and characterize the three segments. First, we quantify the differential effect of trial length on the download and usage behavior of the three segments. Second, we characterize the heterogeneity in the effect of usage on subscription across the three segments. Finally, we correlate a user's optimal treatment with her/his pretreatment demographics and posttreatment outcomes to characterize the three segments. We refer readers to Online Appendix H for the details of these analyses and provide a summary of the three segments below.

- Seven-day optimal segment: A vast majority of these users are beginners or students, and they are the least likely to subscribe. These users use the product more when given longer trials but don't scale up their usage as much as the 14-day optimal segment. This is understandable because most of them lack the skills to

Figure 5. CDF of Estimated CATEs for 7 vs. 30 Days of Free Trial from Using Different Methods (for Test Data)



use the product extensively, even if given the opportunity to do so. Furthermore, the negative effect of long dormancy periods is the most severe for this segment. This is understandable because these users are less familiar with the product, so when they go for a long period without using the software, they are more likely to conclude that the software is not useful and choose not to subscribe.

Overall, we find that short trials are more effective for beginners and new users because even though there are some positive effects of learning and usage, extended periods of inactivity at the end of long trials can have a strong negative effect on their subscription. One might wonder if this result simply stems from the fact that beginners have short tasks that require more than a week to complete (but still less time than 14/30 days), which leads them to have lower subscription rates when assigned the 14- and 30-day trial (i.e., a more complex version of the demand cannibalization hypothesis). However, if this explanation is true, then we should find that beginners/7-day optimal users who are assigned to the 14- and 30-day trial should be less likely to subscribe if they use the product more. However, we find the opposite; see Online Appendix H.4.

- Fourteen-day optimal segment: These users are more likely to be mixed skill, and they have the highest usage and subscription rates. This segment takes the most advantage of longer trials; that is, they use the product the most and have the shortest periods of dormancy when given longer trials. It seems like these users actually try the product's features and evaluate the product carefully before deciding whether to subscribe or not. However, the effect of usage on subscriptions is lower for these users (compared with the other two segments). This is likely because they are figuring out whether the software is a good fit or not, and more usage may lead some users to learn that it is not a good fit. Furthermore, the magnitude of the negative effect of dormancy length on subscription is also high for them. That is why the 30-day trial is not optimal for these users: the higher usage that comes with a more extended trial does not translate to big differences in subscription, but they still get hit by the increase in dormancy length with the 30-day trial. On the other hand, when they are given only 7 days, they cannot use the product much, and the benefit from higher usage is not realized. Thus, 14 days is ideal for these users.
- Thirty-day optimal segment: These users are more experienced than average are less likely to be students and hobbyists, and more likely to sign up through the app manager instead of the website. These factors suggest that these are more likely to be experienced/legacy users who are already familiar with the software. Long dormancy periods have the least negative effect on these users. This is understandable because these users are likely to be already aware of and experienced with the software. Thus, they are unlikely to infer that the product is not useful if they do not use it for a few days at the end of the trial. Furthermore, longer trials lead to more usage for these users, and the effect of usage on subscription is also high. Thus, giving them 30 days for trial is good.

One interesting pattern in the previous findings is the nonmonotonicity of usage across the three segments. We find that 7-day optimal users use the product the least, followed by the 30-day optimal users, whereas the 14-day optimal users use the product the most. This can be explained by the relative expertise levels of the three groups. The extent to which a user uses the product depends on two factors: (1) how much do they need to evaluate the product and (2) how much can they evaluate product? The 7-day optimal users, who are predominantly beginners have the least ability to explore the product features and hence use it the least. In contrast, the 30-day optimal users, who are more likely to be experts and legacy users, have the highest ability to evaluate the product. However, given their expertise and familiarity with the software, they can do this without extensive usage. Finally, the 14-day optimal users, who are more likely to be mixed-skill users, have both high need to

evaluate the product and sufficient ability to explore it. Hence, they have the highest usage.

It worth mentioning that our findings provide partial support to the theories proposed in the literature on the relationship between users' skill-level/experience and the effectiveness of free trials. For example, Dey et al. (2013) argue that longer trials are beneficial only when the learning rate is sufficiently large. We find that this is true in our case as well. However, this prior analytical research does not consider the negative effect of dormancy length on subscription, especially for beginners and new users. They argue that beginners should be given longer free trials because longer trials allow them to learn about the product, which increases their likelihood of subscription. In contrast, we find that short trials are optimal for beginners. Although longer trials have a positive impact on the usage and subscription of this group, they are also the group whose subscription is most negatively affected by longer dormancy periods. Thus, ignoring the negative impact of dormancy length can lead us to make suboptimal allocations of trial lengths for different segments.

Our findings suggest that firms and managers should take into account the heterogeneity in the evolution of usage and inactivity (as trial length increases) for different consumer types and customize trial lengths based on these patterns. In our setting, users require some skill and need to invest the effort to learn and use the software effectively. In particular, beginners and inexperienced users are unable to scale up their usage with longer trials and therefore have longer periods of inactivity later in the trial period (which has a detrimental effect on subscription). However, if the software is simple and easy to use, we would not see such periods of inactivity. Interestingly, this suggests that simpler products may benefit from longer trials (especially for beginners), whereas more complex products may benefit from shorter trials. In sum, both the complexity of the product and the skill of the user jointly determine usage and activity (or inactivity), which then affects subscription. Our results provide some general guidelines to firms on how to pick the right trial length for different products and segments.

7. Long-Term Outcomes: Consumer Loyalty and Profitability

Thus far, we focused on short-run outcomes in our policy design and evaluation. However, a policy that maximizes subscriptions (or short-run conversions) may not be the best long-run policy if it brings in users who are less profitable or less loyal. For example, a policy that increases subscriptions among students (who get a significant educational discount and hence pay lower prices) and/or users who subscribe to lower-end products/bundles (that are priced much lower than the all-inclusive software suite) at the expense of high-end users can lead to lower revenues. Similarly, a policy designed to maximizes subscriptions may do so at the expense of long-term retention; that is, it may bring in the less loyal consumers who churn within a short period. Thus, a subscription-optimal policy may in fact be suboptimal from the perspective of long-run outcomes (Gupta et al. 2006, Fader and Hardie 2009, McCarthy et al. 2017). In this section, we therefore examine two important postsubscription outcomes of interest for the firm.

- Consumer loyalty, as measured by subscription length or the number of months a user subscribes to the service over the two-year period after the experiment.
- Consumer profitability, as measured by the revenue generated by the user over the two years after the experiment. (In SaaS settings, revenues and profits can be treated as equivalent because the marginal cost of serving an additional user is close to zero.)

7.1. Gains in Retention and Revenue from Counterfactual Policies

We first show the average treatment effect of the three trial lengths on retention and revenue in Table 9 and 10.[12] We find that the 7-day trial continues to be the best. In the test data, it increases retention by 6.4% and revenue by 7.91%. The average effect of the 14-day trial is both smaller in magnitude and not significant in the training data.[13] These results largely mirror our findings on the average treatment effect of subscription; that is, the 7-day trial is the best treatment.

Next, we examine how the uniform and personalized targeting policies described in Section 5.2 perform

Table 9. Average Effect of the 7- and 14-Day Treatments on Subscription Length Compared with the Control Condition of 30-Day Free Trial

Data	Treatment	Average subscription length	Retention difference	t statistics	Percentage gain over baseline
Training data	7 days	2.32	0.16	4.27	7.27
	14 days	2.22	0.06	1.54	2.59
	30 days	2.17	—	—	—
Test data	7 days	2.33	0.14	2.42	6.28
	14 days	2.32	0.13	2.27	5.91
	30 days	2.19	—	—	—

Table 10. Average Effect of the 7- and 14-Day Treatments on Revenue Compared with the Control Condition of 30-Day Free Trial

Data	Treatment	Average revenue	Revenue difference	t statistics	Percentage gain over baseline
Training data	7 days	82.65	6.17	3.75	8.06
	14 days	79.08	2.59	1.58	3.38
	30 days	76.49	—	—	—
Test data	7 days	83.72	6.14	2.42	7.92
	14 days	84.02	6.44	2.53	8.30
	30 days	77.58	—	—	—

on the two long-term outcomes of interest. To derive the the IPS estimates of average subscription length and revenue under policy π, we first segment users into three groups based on the policy-prescribed treatment: (1) $\pi(X_i) = 7$ days, (2) $\pi(X_i) = 14$ days, and (3) $\pi(X_i) = 30$ days. Then, we use the observed subscription lengths and revenues as the outcome variables (Y_i) in Equation (4) to estimate the IPS rewards for these outcomes. Table 11 shows the results from this analysis. The main takeaway is that the personalized policy, π_{lasso}, which was designed to maximize subscriptions, also does well on consumer loyalty and revenue compared with the other uniform policies. This is valuable from the the firms' perspective because it suggests that policies optimized for short-run outcomes are largely aligned with long-run outcomes as well.

7.2. Gains in Short-Run vs. Long-Run Outcomes

An interesting empirical pattern here is that the gains in subscription length and revenues are quantitatively different from the gain in subscription (compare the percentage increases in Tables 8 and 11). We now discuss the source of this difference.

We can write down the expected subscription length (denoted by Y_i^l) conditional on treatment W_i as

$$\mathbb{E}(Y_i^l \mid W_i) = \Pr(Y_i^s \mid W_i) \cdot \mathbb{E}\left[T_{end} - T_{start} \mid W_i, Y_i^s = 1\right], \quad (5)$$

where $\Pr(Y_i^s \mid W_i)$ is the probability that user i will subscribe conditional on receiving treatment W_i, and $\mathbb{E}\left[T_{end} - T_{start} \mid W_i, Y_i^s = 1\right]$ is i's expected length of subscription conditional on receiving treatment W_i and subscribing ($Y_i^s = 1$). The reason for the discrepancy in the gains on the two outcomes, subscription and subscription length, becomes apparent from Equation (5). If trial length affects not just subscription, but also how long a subscriber will remain loyal to the firm, then the gains in Y_i^l will be naturally different from the gains in subscription. To examine if this is true in our data, we present the summary statistics for $\mathbb{E}[T_{end} - T_{start} \mid W_i, Y_i^s = 1]$ for the three trial lengths in Table A16 in Online Appendix I. We see that there are some small differences in this metric across the three trial lengths, which account for the differences between the gains in subscription and gains in subscription length.

Similarly, we can write the expected revenue (denoted by Y_i^r) conditional on treatment W_i as

$$\mathbb{E}(Y_i^r \mid W_i) = \Pr(Y_i^s \mid W_i) \cdot \mathbb{E}\left[T_{end} - T_{start} \mid W_i, Y_i^s = 1\right] \cdot \mathbb{E}\left[\text{Price}_i \mid W_i, X_i, Y_i^s = 1\right]. \quad (6)$$

This is similar to Equation (5), with the additional $\mathbb{E}\left[\text{Price}_i \mid W_i, X_i, Y_i^s = 1\right]$ term. It suggests that trial length can influence revenues through three channels: (1) subscriptions, (2) length of subscription, and (3) price of the product subscribed. The first two were already discussed in the previous paragraph. We now examine whether the products that consumers subscribe to and the prices that they pay are also a function of trial length. That is, we examine whether $\mathbb{E}[\text{Price}_i \mid W_i, X_i, Y_i^s = 1]$ is indeed a function of W_i in our data. The price that a subscriber pays is a function of both the product that s/he subscribes to (e.g., single product, all-inclusive bundle) and her/his demographics (e.g., students pay lower prices for the same product). In Table

Table 11. IPS Estimates of the Average Subscription Length and Revenue Under Counterfactual Policies (Three Uniform and One Personalized)

		Subscription length				Revenue			
		Estimate (mo)		Increase (%)		Estimate ($)		Increase (%)	
Category	Policy	Training	Test	Training	Test	Training	Test	Training	Test
Personalized	π_{lasso}	2.39	2.36	10.42	7.96	85.96	86.67	12.38	11.72
Uniform	π_7	2.32	2.33	7.27	6.28	82.65	83.72	8.06	7.92
	π_{14}	2.22	2.32	2.59	5.91	79.08	84.02	3.38	8.30
	π_{30} (Baseline)	2.17	2.19	—	—	76.49	77.58	—	—

Table 12. Expected Mean of the Three Outcomes of Interest Under Policies Optimizing Each Outcome

Data set	Policy optimized on	Subscription	Total revenue	Subscription length
Training data	Subscription	15.85	85.96	2.39
	Total Revenue	15.60	86.41	2.35
	Subscription Length	15.71	84.77	2.40
Test data	Subscription	15.62	86.67	2.36
	Total Revenue	15.45	84.28	2.33
	Subscription Length	15.53	84.86	2.35

A17 in Online Appendix I, we present the distribution of products and subscription type by trial length for all the subscribers in our data. Again, we see that there are some minor differences in product and subscription types across trial lengths, which explain the differences in revenue gains.

7.3. Optimizing on Long-Run Outcomes

Thus far, we saw that a personalized policy designed to optimize short-run conversions also does well on long-run outcomes. However, this still begs the question of how it compares to policies directly optimized to maximize long-run outcomes. In practice, the problem with using retention/revenues until some period T (e.g., two years) is that we have to wait until T to identify the best policy and then implement it. This is both suboptimal and impractical from a firm's perspective. In contrast, using a short-term outcome such as subscriptions to design policy and then projecting the policy gains on long-term objectives (e.g., revenues) is both practical and feasible. However, a policy optimized on short-run outcomes may still perform worse than one directly optimized on long-term outcomes. Therefore, we examine and compare the performance of the policy designed to maximize subscriptions with policies designed to maximize customer loyalty or profitability and see which performs better in our context.

To that end, we now design two other personalized policies designed to maximize: (1) subscription length and (2) revenue. The policy design follows the same procedure described in Section 5.1, but with revenue (Y_i^r) and subscription length (Y_i^l) as our outcome variables. That is, we first estimate two separate lasso models with the previous two variables as outcome variables and then assign policy based on them.

Table 12 compares the performance of the three personalized policies on the three outcome variables of interest to the firm: subscriptions, subscription length, and revenue.[14] Interestingly, we find that the policy optimized on short-run conversions (subscriptions) also performs the best on retention and revenue. There are three reasons for this. First, as we saw in the previous section, conditional on subscription, the differences in retention length and products purchased are relatively minor. Hence, optimizing on subscription is largely consistent with optimizing on retention/revenue. Second, recall that the long-run outcomes are missing for about 7% of the subscribers. Therefore, the policies based on these outcomes have less information for training, which compromises their generalizability and hampers their performance on the test data. Third, subscription is a binary outcome and as such has no variance in the positive realizations. In contrast, the variance in the long-run outcome variables (subscription length and revenue) can be quite high. This variance makes it harder to generalize models based on these outcomes, which in turn adversely affects the performance of policies based on them.

In summary, our findings suggest that if there are no significant differences in customer loyalty and profitability as a function of the promotional channel through which the user converted (trial length in this case), then optimizing low-variance short-run conversions will also lead to more generalizable policies that will also perform well on long-run outcomes.

8. Conclusions

Free trials are now a commonly used promotional strategy for SaaS products and other digital experience goods. In this paper, we examine the effect of trial length on consumers' subscription and retention decisions using data from a large-scale field experiment run by a leading SaaS firm, where the firm randomly assigned new users to 7, 14, or 30 days of free trial. We leverage two unique features of the data in our study: (1) the exogenous assignment of trial length and (2) the user's posttreatment product download and usage data during the trial period.

We find that the shortest trial length (7 days) is the best average treatment and maximizes both short- and long-run outcomes, customer acquisition, retention, and profitability. Although this result is likely to be specific to our setting, we examine the behavioral underpinning of these findings and provide some evidence on the mechanisms at play. We rule out the demand cannibalization or free riding theory, find support for the consumer learning hypothesis, and identify a novel mechanism that plays a significant role in the effectiveness of free trials: the negative effect of long stretches of inactivity at the end of the trial on subscription.

We then develop a personalized targeting policy based on lasso and show that it can lead to a more than 6.8% improvement in subscription compared with the baseline uniform policy of giving a 30-day trial for all. Further exploration of usage within different consumer segments in our personalization scheme suggests that simpler products and experienced users are more likely to benefit from longer trials. Finally, we find that the personalized policy designed to optimize subscriptions also performs well on long-term metrics such as customer retention and revenues in our setting. We also compare the performance of our benchmark personalized policy with alternative personalized policies developed based on other well-known outcome estimators (e.g., random forests) and the recently developed heterogeneous treatment effects estimators (e.g., generalized random forests). We find that many alternative personalized policies perform poorly and are often worse than the simple uniform 7 days for all policy. A key managerial takeaway is that firms should not naively assume that personalization based on the most advanced estimators always helps. Instead, they should develop personalized policies based on a number of methods and carefully evaluate them using offline IPS estimators before investing resources in deploying personalized policies in the field.

Our paper opens many avenues for future research. First, although our analysis indicates that product usage during the trial period affects users' subscription decisions, we do not causally tie usage to subscriptions because usage is endogenous. Nevertheless, future research may be able to use treatment assignment (e.g., trial length) as an instrument that exogenously shifts usage and directly estimate the effect of usage on purchase. This can provide further insight into the question of whether encouraging usage (either through free trial or product improvements) can lead to better purchase outcomes. Second, our analysis suggests that personalized policies do not always perform better than a simple uniform policy. One interesting finding is that outcome estimators that have high predictive ability do not necessarily do well on personalized policy design (compare the performance of models in Table A9 in Online Appendix G with Table 8). Moreover, the finding that recently developed CATE estimators such as causal forest do not perform well in our setting is also surprising. Further investigation into the question of whether these results are generalizable would be a useful next step.

Acknowledgments

The authors thank an anonymous firm for providing the data; Atanu Lahiri, Simha Mummalaneni, and Omid Rafieian for detailed comments that improved the paper; and the participants of the 2018 Marketing Science conference and the Triennial Choice Symposium, 2021 UT Dallas Bass Conference and seminar audiences at the Emory University, Johns Hopkins University, Lehigh University, University of California at Berkeley, University of Houston, University of Maryland, University of Rochester, University of South Carolina, and University of Southern California for their feedback.

Endnotes

[1] This free trial is at the software product level, that is, users start a separate trial for each software product, and their trial for a given product expires 30 days from the point at which they started the free trial for it.

[2] Although the distribution of prices shown to users is the same across all treatment arms, the distribution of prices paid by the subscribers can be different under each treatment arm. This is because the length of the free trial can influence which types of consumers subscribe and the products/bundles that they subscribe to. These differences can lead to downstream differences in the revenues under treatments and targeting policies. We discuss these issues in detail in Section 7.

[3] Only a small fraction of people choose to not report these data. For example, the percentage of users with "unknown" Skill and Job is 7.4% and 21.9%, respectively.

[4] If a user unsubscribes for a period of time and then comes back, her subscription length is the total number of months that she was a paying customer of the firm. If a user subscribes to two or more plans, we aggregate the length of subscription all plans and report the total. Therefore, the subscription length can be greater than 24 months for such users.

[5] In general, it is a better practice to obtain ATEs directly based on mean comparisons without using regression-based approaches (Imbens and Rubin 2015). Nevertheless, in Online Appendix B.1, we present the ATEs based on regressions (with and without controls), and they are statistically indistinguishable from those shown in the main text.

[6] Furthermore, presenting tables of means for each pretreatment variable is not feasible in our case because all our pretreatment variables are categorical with a large number of subcategories.

[7] Later in the paper, we use the empirical propensities to evaluate the gains from our models. Therefore, any minor discrepancies in the propensities of treatment allocation are taken care of; see Equation (4) and the discussion around it.

[8] The results in Table 7 should only be interpreted as suggestive evidence for the second half of the mechanism shown in Figure 3 (and not causally). This is because the unobserved attributes of the user that drive use may also drive subscription.

[9] The total number of possible policies is $W^{\prod_{d=1}^{D} c_d}$, when we have D pretreatment variables and the dth variable can take c_d different values. This number can be extremely high even in simple settings. In our application, the cardinality of the policy space is equal to $3^{987,840}$.

[10] In theory, in a randomized experiment, the propensity of treatment assignment is orthogonal to the treatment prescribed by any policy π. Thus, $e(W_i = w, X_i) = e(W_i = w)\ \forall w \in \mathcal{W}$ is known and constant for all observations. However, in practice, within the set of users for whom policy π prescribes w, the empirical treatment propensities might not be the same as that in the full data. Therefore, we use the empirical propensity, defined as $\hat{e}_{\pi(X_i)}(W_i) = \frac{\frac{1}{N}\sum_{j=1}^{N} 1_{[W_j = W_i, \pi(X_j)=\pi(X_i)]}}{\frac{1}{N}\sum_{j=1}^{N} 1_{[\pi(X_j)=\pi(X_i)]}}$.

[11] For the outcome estimators, we can estimate treatment effects from outcome estimates as $\mathbb{E}[Y \mid X_i = x, W_i = 7\text{days}] - \mathbb{E}[Y \mid X_i = x, W_i = 30\text{days}]$. For heterogeneous treatment methods, these estimates are directly available (see Equation (A.4) in the online appendix).

[12] A minor point is that we do not have access to subscription length and revenue data for all subscribers (recall the discussion in Section 3.3.2). Therefore, we treat the missing observations as zero in calculations. The results remain qualitatively unchanged if we instead work with the subset of users for whom we have nonmissing revenue data.

[13] Note that the 14-day trial outperforms the 7-day trial in the test data (on revenue) even though the 7-day trial is the best policy in the training data. We present a brief explanation for this discrepancy now. In general, estimates from one data set are valid in another data set only when the joint distribution of outcomes and covariates are similar in both data sets. However, in finite samples, there are usually some minor differences in training and test data because of the randomness in the splitting procedure. In our case, the main difference is this: The distributions of subscription length for the 14-day condition in the training and test data are different. This is, however, not the case for the 7- or 30-day conditions; see Table A16 in Online Appendix I. Thus, the estimate of subscription length from the training data does not translate well to test data, and this leads to the large difference in the subscription length and revenue estimates across the training and test data sets.

[14] See Online Appendix I for a discussion of how the treatment allocation varies across the three policies.

References

Anderson ET, Simester DI (2004) Long-run effects of promotion depth on new vs. established customers: Three field studies. *Marketing Sci.* 23(1):4–20.

Athey S, Wager S (2020) Efficient policy learning. Preprint, submitted September 4, https://arxiv.org/abs/1702.02896.

Bruhn M, McKenzie D (2009) In pursuit of balance: Randomization in practice in development field experiments. *Amer. Econom. J. Appl. Econom.* 1(4):200–232.

Cheng HK, Liu Y (2012) Optimal software free trial strategy: The impact of network externalities and consumer uncertainty. *Inform. Systems Res.* 23(2):488–504.

Dey D, Lahiri A, Liu D (2013) Consumer learning and time-locked trials of software products. *J. Management Inform. Systems* 30(2):239–268.

Dudík M, Langford J, Li L (2011) Doubly robust policy evaluation and learning. *Proc. 28th Internat. Conf. on Machine Learn.* (Omnipress), 1097–1104.

Fader PS, Hardie BG (2009) Probability models for customer-base analysis. *J. Interactive Marketing* 23(1):61–69.

Fong N, Zhang Y, Luo X, Wang X (2019) Targeted promotions on an e-book platform: Crowding out, heterogeneity, and opportunity costs. *J. Marketing Res.* 56(2):310–323.

Foubert B, Gijsbrechts E (2016) Try it, you'll like it—or will you? The perils of early free-trial promotions for high-tech service adoption. *Marketing Sci.* 35(5):810–826.

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* 33(1):1.

Gartner (2019) Forecast: Public cloud services, worldwide, 2016-2022, 4q18 update. Accessed August 1, 2022, https://www.gartner.com/en/newsroom/press-releases/2019-04-02-gartner-forecasts-worldwide-public-cloud-revenue-to-g.

Guo T, Sriram S, Manchanda P (2021) The effect of information disclosure on industry payments to physicians. *J. Marketing Res.* 58(1):115–140.

Gupta S, Hanssens D, Hardie B, Kahn W, Kumar V, Lin N, Ravishanker N, Sriram S (2006) Modeling customer lifetime value. *J. Serv. Res.* 9(2):139–155.

Hauser JR, Urban GL, Liberali G, Braun M (2009) Website morphing. *Marketing Sci.* 28(2):202–223.

Hitsch GJ, Misra S (2018) Heterogeneous treatment effects and optimal targeting policy evaluation. Preprint, submitted February 6, https://dx.doi.org/10.2139/ssrn.3111957.

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* 47(260):663–685.

Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, Cambridge, UK).

Kitagawa T, Tetenov A (2018) Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86(2):591–616.

Lewis RA, Rao JM (2015) The unfavorable economics of measuring the returns to advertising. *Quart. J. Econom.* 130(4):1941–1973.

Manski CF (2004) Statistical treatment rules for heterogeneous populations. *Econometrica* 72(4):1221–1246.

McCarthy DM, Fader PS, Hardie BG (2017) Valuing subscription-based businesses using publicly disclosed customer data. *J. Marketing* 81(1):17–35.

McKenzie D (2017) Should we require balance t-tests of baseline observables in randomized experiments? Accessed August 1, 2022, https://blogs.worldbank.org/impactevaluations/should-we-require-balance-t-tests-baseline-observables-randomized-experiments.

Mela CF, Gupta S, Lehmann DR (1997) The long-term impact of promotion and advertising on consumer brand choice. *J. Marketing Res.* 34(2):248–261.

Mutz DC, Pemantle R, Pham P (2019) The perils of balance testing in experimental design: Messy analyses of clean data. *Amer. Statist.* 73(1):32–42.

Pauwels K, Hanssens DM, Siddarth S (2002) The long-term effects of price promotions on category incidence, brand choice, and purchase quantity. *J. Marketing Res.* 39(4):421–439.

Prentice RL (1989) Surrogate endpoints in clinical trials: Definition and operational criteria. *Statist. Medicine* 8(4):431–440.

Rafieian O (2019a) Optimizing user engagement through adaptive ad sequencing. Technical report, Cornell Tech, New York.

Rafieian O (2019b) Revenue-optimal dynamic auctions for adaptive ad sequencing. Technical report, Cornell Tech, New York.

Rafieian O, Yoganarasimhan H (2021) Targeting and privacy in mobile advertising. *Marketing Sci.* 40(2):193–218.

Schwartz EM, Bradlow ET, Fader PS (2017) Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Sci.* 36(4):500–522.

Scott CA (1976) The effects of trial and incentives on repeat purchase behavior. *J. Marketing Res.* 13(3):263–269.

Simester D, Timoshenko A, Zoumpoulis SI (2020a) Efficiently evaluating targeting policies: Improving on champion vs. challenger-experiments. *Management Sci.* 66(8):3412–3424.

Simester D, Timoshenko A, Zoumpoulis SI (2020b) Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Sci.* 66(6):2495–2522.

Sunada T (2018) Customer learning and revenue-maximizing trial design. Technical report, University of Rochester, Rochester, NY.

Swaminathan A, Joachims T (2015) Counterfactual risk minimization: Learning from logged bandit feedback. Bach F, Blei D, eds. *Proc. Internat. Conf. Machine Learn.*, vol. 37, 814–823.

Swaminathan A, Krishnamurthy A, Agarwal A, Dudik M, Langford J, Jose D, Zitouni I (2017) Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 3632–3642.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. B* 58(1):267–288.

VanderWeele TJ (2013) Surrogate measures and consistent surrogates. *Biometrics* 69(3):561–565.

Wang S, Özkan-Seely GF (2018) Signaling product quality through a trial period. *Oper. Res.* 66(2):301–312.

Yang J, Eckles D, Dhillon P, Aral S (2022) Targeting for long-term outcomes. Preprint, submitted April 9, https://arxiv.org/abs/2010.15835.

Yoganarasimhan H (2020) Search personalization using machine learning. *Management Sci.* 66(3):1045–1070.

Zhu M, Yang Y, Hsee CK (2018) The mere urgency effect. *J. Consumer Res.* 45(3):673–690.