Efficiently Evaluating Targeting Policies:
Improving Upon Champion vs. Challenger Experiments

March 2019

Duncan Simester
MIT

Artem Timoshenko
MIT

Spyros I. Zoumpoulis
INSEAD

Champion versus challenger field experiments are widely used to compare the performance of different targeting policies. These experiments randomly assign customers to receive marketing actions recommended by either the existing (champion) policy or the new (challenger) policy, and then compare the aggregate outcomes. We propose an alternative experimental design and an alternative estimation approach to improve the evaluation of targeting policies.

Our experimental design randomly assigns customers to marketing actions. This allows evaluation of any targeting policy without requiring an additional experiment, including policies designed after the experiment is implemented. The proposed estimation approach identifies customers for whom different policies recommend the same action and recognizes that for these customers there is no difference in performance. This allows for a more precise comparison of the policies.

We illustrate the advantages of the proposed experimental design and the proposed estimation approach using data from an actual field experiment. We also demonstrate that the grouping of customers, which is the foundation of our estimation approach, can help to improve the training of new targeting policies.

Previous versions of this paper circulated under the titles: “Efficiently Evaluating Targeting Policies Using Field Experiments” (August 2017) and “Evaluating and Improving Targeting Policies with Field Experiments Using Counterfactual Policy Logging” (September 2018).

1. Introduction

Targeting policies are used in marketing to match different marketing actions to different customers. For example, retailers want to send different promotions to different customers, media owners want to show different digital advertisements to different users, online entertainment platforms recommend different content to different customers, real estate agents want to show different homes, financial advisors want to recommend different products, and car dealers want to propose different prices.

A standard approach to measuring whether a new policy outperforms an existing policy is to conduct a "champion vs. challenger" field experiment. The new (challenger) policy is used to choose marketing actions for a randomly selected group of participants, while a second randomly selected group receives the marketing actions recommended by the existing (champion) policy. The participants' responses are then typically used to calculate the aggregate outcome for each policy, and these aggregate outcomes are compared across policies. For example, Skiera and Nabout (2013) propose a model that targets different search engine keywords with different bids by an advertiser. They test their model using a field experiment in which bids for twenty keywords were submitted using either the current policy or the proposed model. Similarly, Mantrala et al. (2006) proposed a model for setting different prices for different automobile parts. They validated their model using a field experiment in which 200 stores were randomly assigned to the proposed policy, and 300 stores were randomly assigned to the current policy.

We propose using both a different experimental design and a different estimation approach. We start by describing the proposed experimental design.

Proposed Experimental Design for Evaluating Targeting Policies

A targeting policy assigns marketing actions (such as promotions) to different customers. A standard champion vs. challenger design uses a randomized-by-policy (RBP) approach, in which customers are randomly assigned to targeting policies. Instead, we recommend a randomized-by-action (RBA) design, in which customers are randomly assigned to marketing actions. For example, consider a problem of recommending chocolate, strawberry, or vanilla ice cream flavor to customers. An RBP experimental design would use one targeting policy to assign flavors to customers in one experimental condition, and the alternative targeting policy to assign flavors to different customers in another condition. In contrast, an RBA design randomly assigns chocolate, strawberry, and vanilla to different groups of customers.

A disadvantage of the RBP design is that it will often not be possible to evaluate new targeting policies using data from an RBP design. For example, if the RBP design implemented policies that only recommended chocolate and strawberry, we could not evaluate a new policy that recommended vanilla. The RBA design allows the evaluation of new policies without an additional experiment. Because the

RBA design randomly assigns flavors to customers, whenever a new policy recommends vanilla to some customers, some of them will have actually received vanilla (and similarly for the other flavors).[1] The customers that received the flavors recommended for them can be used to evaluate the new policy.

We next discuss the proposed estimation procedure, which can use data from an RBA or RBP experiment.

Proposed Estimation Approach for Comparing Targeting Policies

The standard approach for comparing targeting policies estimates the aggregate performance of each policy separately and then calculates the difference in the aggregate performances. Instead, we propose an approach that identifies customers for whom different policies recommend the same action and recognizes that for these customers the performance is identical.

For example, consider two targeting policies that both recommend chocolate for male customers. For female customers, assume they recommend different actions; the first policy recommends strawberry and the second policy recommends vanilla. To compare these policies we construct two groups of customers: customers for whom the policies recommended the same actions (male customers), and customers for whom they recommended different actions (female customers). For males we know that the true difference in the performance of the two policies is exactly zero (because both policies recommend chocolate). For females we just need to compare the outcomes for strawberry versus vanilla. This comparison is possible in both RBA and RBP experimental designs. In the RBA design, the flavor assignment is random, so there is a random group of females that actually received strawberry, and a random group of females that received vanilla. In the RBP design, a random group of females received actions recommended by the first policy (strawberry), and another random group received actions recommended by the second policy (vanilla). Because the assignments are random, we can safely compare the outcomes between the groups without concern for customer differences.

The advantage of the proposed estimation approach is that it improves precision when comparing the performance of two policies. Recognizing that the performance is identical when two policies recommend the same action removes random error due to differences in observed performances.

Training New Policies

The data from an RBA experiment can be used to train and evaluate new policies. However, training a new policy solely using the new data from an RBA experiment may mean losing some of the information

[1] Exceptions may arise if a marketing action is very rarely recommended by a targeting policy. For example, if a targeting policy recommends sending vanilla to just one customer in the population, an RBA design may not assign vanilla to that customer. These rare exceptions have little practical importance because they will have almost no impact on the overall performance of the policy. The Horvitz-Thompson estimator could be used for policy evaluation when these exceptions arise (Horvitz and Thompson 1952).

that was used to create existing policies. Our estimation approach suggests a way to retain that information. The cornerstone of the estimation approach is the grouping of customers using the recommendations from targeting policies. This grouping provides a convenient way to summarize the information contained in existing policies. We use the data from an actual field experiment to demonstrate that logging actions recommended by the existing policies helps to preserve and incorporate this information when training a new policy, leading to improvement in the performance of the new policy.

Applications

To illustrate potential applications of the proposed approach, we highlight how the proposed approach can contribute to three recent papers that study targeting of marketing actions. Dubé and Misra (2017) train a price targeting model using data from one experiment and then validate this model using a second experiment. The design of their first (training) experiment represents an RBA design, where the authors randomly assign marketing actions to the customers. The design of their second (validation) experiment represents an RBP design, and includes two uniform benchmark policies and their proposed policy. A challenge to using an RBA design in their second experiment is that the range of prices a firm can charge is continuous, and so the action space is infinite. One solution is to discretize this continuous variable. For example, Dubé and Misra (2017) round targeted prices down to the nearest $9 price ending, yielding 39 possible prices ranging from $119 to $499 in $10 increments. Even without changing the experimental design, the proposed estimation approach could improve the efficiency of their comparisons. For subjects for whom their optimized policy recommends the same price as one of the benchmark policies, we know that the true performance difference between the proposed policy and the benchmark policy is zero. Setting this difference to zero would eliminate variance introduced by random noise.

Ostrovsky and Schwarz (2011) study how to set reserve prices in Internet advertising auctions. They present the results of an RBP experiment in which reserve prices were randomly assigned either to a uniform benchmark price or to a price proposed by the targeting model. Similar to Dubé and Misra (2017), the RBA design would require discretizing prices, but would enable evaluation of a much wider range of targeting policies. Our proposed estimation approach would also improve the efficiency of their comparisons of the candidate policies.

Rafieian and Yoganarasimhan (2018) present an example of a targeting problem with a large action space, where an RBA experimental design is actually implemented. This paper studies targeting of mobile advertising at a large advertisement network. The platform uses a quasi-proportional auction mechanism to allocate advertisement positions, which ensures positive probabilities of displaying advertisements by all bidders. This is consistent with the RBA design. Rafieian and Yoganarasimhan (2018) compare their proposed model with the firm's current model. Grouping customers according to whether different

policies recommend the same actions or different actions has the potential to improve the efficiency of this performance comparison. It also provides a convenient way to use the information in the current policy when training a new targeting policy.

Outline of the Paper

The paper continues in Section 2 with a review of the literature. We illustrate the efficiency improvements using a formal model in Section 3. In Section 4 we describe how to use OLS to compare the performance of two policies. In Section 5 we present an empirical application that we use to highlight the benefits of the method. In Section 6 we discuss how to incorporate information from existing policies when training new targeting policies using RBA data. Limitations are highlighted in Section 7 and the paper concludes in Section 8.

2. Literature Review

The contrast between an RBP (randomization-by-policy) and an RBA (randomization-by-action) design has its roots in the reinforcement learning literature distinguishing on-policy and off-policy evaluation (Sutton and Barto, 1998). In on-policy evaluation, the evaluation data is constructed by implementing the policy. In off-policy evaluation, the policy is evaluated using data constructed by implementing a different policy (Langford et al., 2008; Strehl et al., 2010; Dudík et al., 2011). An RBP experiment is an example of on-policy evaluation, while an RBA design is an example of off-policy evaluation.

The paper is related to research in marketing and other fields that has focused on failure of an intention to treat. The failure of an intention to treat typically reflects a lack of compliance. For example, subjects may not open promotional mail sent to them, may not take their assigned medicine, or may not complete their assigned education. As a result, some customers in the Treatment group receive essentially the same treatment as subjects in a Control group. In our setting, when comparing targeting policies there is no difference in treatments when two targeting policies recommend the same action. This insight is a central feature of this paper; it forms the basis of the efficiency improvements that we highlight.

Within the marketing literature there has been a growth of interest in using experiments to evaluate targeting policies. Our work is most closely related to three studies. In a recent working paper, whose research coincides with our own research on this topic, Hitsch and Misra (2018) highlight the advantages of randomly assigning customers to marketing actions. They recognize that data from an RBA experiment can be used to evaluate any targeting policy. However, they focus their discussion on the distinction and comparison between direct and indirect methods of estimating conditional average treatment effects, which is a question that lies outside of our scope.

Johnson, Lewis and Reiley (2017) propose an estimation approach that has similar features to the estimation approach we recommend. They recognize that an intent to treat on Yahoo! may fail if: "many users in the experiment *do not see an ad* because they either do not visit Yahoo! at all or do not browse enough pages on Yahoo! during the campaigns". The true difference in the treatment and control outcomes for these customers is zero. They identify these customers in both the treatment and control conditions and then remove them from the estimation sample. This is similar to our insight that there is no difference in the performance of two targeting policies among customers for whom they recommend the same action. However, comparing targeting policies is a different problem from estimating the average effect of an advertising treatment on the treated (TOT). As we will discuss, simply omitting customers for whom the true difference in performance is zero would distort estimates of the average difference in the performance of the candidate policies.

Johnson, Lewis and Nubbemeyer (2017) provide a similar example. They recognize that a challenge in measuring the effectiveness of online advertising is that advertising platforms allocate exposures to customers systematically. As a result, the customers who see an advertisement are different than customers who do not, and we cannot simply compare purchasing behavior of these two groups. In contrast, comparing all of the customers that the platform intends to treat with an equivalent group of customers that the platform does not intend to treat is inefficient, because the outcomes for customers who will never be treated just add noise. Removing customers who would never be treated from both the treatment and the control group allows for a more precise comparison between the two groups.

We present a formal model of the proposed approach in the next section. The model illustrates the efficiency benefits of the proposed approach and motivates the estimation section that follows.

3. Model

We consider customers $h = 1, \dots, H$ from a population $\mathcal{H}$. For each customer h, there is a vector of observable covariates, $\boldsymbol{x}_h \in \mathcal{X}$. The firm chooses which marketing action each customer will receive.[2] We assume that the set of marketing actions $\mathcal{A}$ is finite. For each customer h and marketing action $a \in \mathcal{A}$, we define the monetary outcome $Y_h(a)$ if customer h is treated with marketing action a. The outcome $Y_h(a)$ is a random variable:

$$Y_h(a) = \alpha_a + \beta \boldsymbol{x}_h + \varepsilon_{a,h}$$

We assume that the marketing actions have an additive constant effect on the outcome, and the random error terms, $\varepsilon_{a,h}$, are i.i.d. with $\mathbb{E}[\varepsilon_{a,h}] = 0$.

[2] This assumption may not always hold. For example, in a digital advertising setting, the advertising platform may make it difficult for a firm to control which advertisement a customer will receive.

We define a targeting policy $\mathcal{P}$ as a function $\mathcal{P}(h)\colon \mathcal{H} \to \mathcal{A}$.[3] We will consider a set of $T \geq 2$ targeting policies that the firm wants to test, which we denote $\mathcal{P}_1, \ldots, \mathcal{P}_T$. We define the value of a targeting policy $\mathcal{P}$ to be the measure:

$$V(\mathcal{P}) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{E}[Y_h; \boldsymbol{x}_h, \mathcal{P}(h)].$$

Experiments

We assume the firm implements a validation experiment with L experimental conditions, indexed $1, \ldots, L$. Each customer h is randomly assigned to one of L experimental conditions; we denote the assignment of customer h by W_h. We assume that the assignment to the experimental conditions is independent of the customer covariates x_h.

In a randomized-by-policy (RBP) design, each experimental condition corresponds to a targeting policy. A customer h in experimental condition $W_h = \mathcal{P}_w$ is assigned to receive marketing action $\mathcal{P}_w(h)$, which is the action that policy $\mathcal{P}_w$ recommends for her. The number of experimental conditions in this experimental design is equal to the number of candidate targeting policies, $L = T$. We define Y_h^{obs} as the observed outcome of customer h in the experiment. Given $W_h = \mathcal{P}_w$, the outcome Y_h^{obs} is a realization of the random variable $Y_h(a)$, where $a = \mathcal{P}_w(h)$ is the action that policy $\mathcal{P}_w$ recommends for customer h. We assume that the outcome $Y_h(a)$ depends on the recommended marketing action $a = \mathcal{P}_w(h)$, and not on the targeting policy $\mathcal{P}_w$ per se.[4]

In a randomized-by-action (RBA) design, each experimental condition corresponds to a marketing action. A customer in an experimental condition is assigned the marketing action that corresponds to that condition. We write $W_h = a_w$ to denote that customer h is in the experimental condition that receives marketing action a_w. The number of experimental conditions in this experimental design is equal to the number of marketing actions that the firm is considering, $L = |\mathcal{A}|$. Given $W_h = a_w$, the observed outcome Y_h^{obs} is a realization of random variable $Y_h(a)$, although now $a = a_w$ rather than $a = \mathcal{P}_w(h)$.

Evaluating a Single Policy

Consider a single targeting policy $\mathcal{P}_1$. When evaluating $\mathcal{P}_1$ our goal is to estimate $V(\mathcal{P}_1)$. We use the notation $\hat{V}(\mathcal{P}_1)$ to denote an estimator of $V(\mathcal{P}_1)$.

[3] A targeting policy can also be defined as a function $\mathcal{P}(\boldsymbol{x}_h)\colon \mathcal{X} \to \mathcal{A}$ from customers' covariates to actions.
[4] This assumption is likely to hold in most targeting applications. However, there are settings in which the assumption may not hold. For example, in a two-sided market, a targeting policy may itself have an effect on the outcome, through its effect on the market equilibrium.

We begin with the construction of a sample of observations. We will label the *Policy* $\mathcal{P}_1$ *Dataset* as the set of observations (customers) that were randomly assigned to receive the treatment recommended by $\mathcal{P}_1$. Using an RBP experiment we define:

$$Policy\ \mathcal{P}_1\ Dataset_{RBP} \equiv \{h: W_h = \mathcal{P}_1\},$$

and in an RBA experiment we define:

$$Policy\ \mathcal{P}_1\ Dataset_{RBA} \equiv \{h: \mathcal{P}_1(h) = W_h\}.$$

To evaluate policy $\mathcal{P}_1$ using an RBB experiment or an RBA experiment, we use the respective dataset to calculate a simple mean:

$$\hat{V}(\mathcal{P}_1) = \frac{1}{|Policy\ \mathcal{P}_1\ Dataset|} \sum_{h \in Policy\ \mathcal{P}_1\ Dataset} Y_h^{obs},$$

where the $Policy\ \mathcal{P}_1\ Dataset$ is the $Policy\ \mathcal{P}_1\ Dataset_{RBP}$ or the $Policy\ \mathcal{P}_1\ Dataset_{RBA}$, respectively.

With an RBP design, we need as many experimental conditions as the targeting policies we are testing. To test a new targeting policy, we generally need to implement a new experimental condition. In contrast, data from an RBA design allow the evaluation of any targeting policies, including policies designed after the experiment is conducted (subject to the rare exceptions mentioned in footnote 1). The proposed RBA design is also economical when evaluating targeting policies that assign actions from a small action space; the number of experimental conditions is only as large as the action space.

Comparing Two Policies

Now consider two targeting policies $\mathcal{P}_1$ and $\mathcal{P}_2$. The traditional approach to comparing the performance of two targeting policies is to evaluate the value of each policy separately, $V(\mathcal{P}_1)$ and $V(\mathcal{P}_2)$, and then to calculate the difference:

$$V(\mathcal{P}_1) - V(\mathcal{P}_2) = \frac{1}{H}\sum_{h=1}^{H} \mathbb{E}[Y_h; x_h, \mathcal{P}_1(h)] - \frac{1}{H}\sum_{h=1}^{H} \mathbb{E}[Y_h; x_h, \mathcal{P}_2(h)].$$

Instead of evaluating the targeting policies separately, we propose an alternative approach that splits the population of customers $\mathcal{H}$ into two groups. The first group of customers includes customers for whom policies $\mathcal{P}_1$ and $\mathcal{P}_2$ recommend the same marketing action: $\{h: \mathcal{P}_1(h) = \mathcal{P}_2(h)\}$. For this group, the true difference in performance between the two policies is exactly zero. The second group of customers includes customers for whom the policies recommend different actions: $\{h: \mathcal{P}_1(h) \neq \mathcal{P}_2(h)\}$. We propose comparing the performance of $\mathcal{P}_1$ and $\mathcal{P}_2$ by focusing directly on this second group of customers:

$$V(\mathcal{P}_1 - P_2) = \frac{1}{H}\sum_{h:\,\mathcal{P}_1(h)\neq\mathcal{P}_2(h)} \{\mathbb{E}[Y_h; x_h, \mathcal{P}_1(h)] - \mathbb{E}[Y_h; x_h, \mathcal{P}_2(h)]\}$$

The difference between the traditional approach and the proposed policy comparison approach is the treatment of the first group of customers, for which the two policies recommend the same action. In the traditional approach we use the observed outcomes for these customers. In the proposed approach we set both the performance difference and the variance of the difference to zero. Because of random noise, the observed performance differences may not equal zero, and so the proposed approach removes a source of random error and provides a more efficient comparison. This is particularly relevant when using RBP data, but is also relevant when using RBA data. In Section 5 we will illustrate this efficiency advantage using both types of data. In the next section we describe how to use OLS to estimate the difference in performance between two policies.

4. Using OLS to Compare the Performance of Two Policies

We describe how to use OLS and data from either an RBP or RBA experiment to compare two targeting policies $\mathcal{P}_1$ and $\mathcal{P}_2$.[5] We start by describing the traditional approach to comparing two policies using OLS.

Traditional Approach

The traditional approach is to use data from an RBP experiment and estimate the following model:

$$Y_h = \alpha + \gamma \cdot \ Policy\ 1_h + \boldsymbol{\beta}\boldsymbol{x}_h + \ \varepsilon_h. \qquad (1)$$

The estimation sample includes all of the observations (customers) in the two RBP experimental conditions associated with $\mathcal{P}_1$ and $\mathcal{P}_2$. This is the set of customers:

$$Policy\ \mathcal{P}_1\ Dataset_{RBP} \cup Policy\ \mathcal{P}_2\ Dataset_{RBP} = \{h: W_h \in \{\mathcal{P}_1,\ \mathcal{P}_2\}\}.$$

The $Policy\ 1$ variable is a binary indicator identifying whether customer h was in the experimental condition associated with $\mathcal{P}_1$, and $\boldsymbol{x}_h$is a vector of covariates. The coefficient of interest is γ, which represents an average treatment effect measuring the increase (or decrease) in the observed outcome Y_h. Equation 1 also includes covariates, which is a standard approach for improving the efficiency of the estimate of γ.

[5] A direct estimation approach is also possible. However, direct estimation cannot easily incorporate covariates, and so we relegate this to Appendix A.

This approach yields an unbiased estimate of the difference in the performance of $\mathcal{P}_1$ and $\mathcal{P}_2$. However, the estimate is not as efficient as the proposed estimation approach. We describe the proposed estimation approach using RBP data next.

Proposed Estimation Using RBP Data

We divide the customers in the two conditions associated with policies $\mathcal{P}_1$ and $\mathcal{P}_2$ into two groups:

> **Same Recommendations Group (RBP)**: the set of customers in the union of the $Policy\ 1\ Dataset_{RBP}$ and the $Policy\ 2\ Dataset_{RBP}$ for whom both policies recommend the same action:
>
> $$\{h: W_h \in \{\mathcal{P}_1, \mathcal{P}_2\}, \mathcal{P}_1(h) = \mathcal{P}_2(h)\}.$$
>
> **Different Recommendations Group (RBP)**: the set of customers in the union of the $Policy\ 1\ Dataset_{RBP}$ and the $Policy\ 2\ Dataset_{RBP}$ for whom the two policies recommend different actions:
>
> $$\{h: W_h \in \{\mathcal{P}_1, \mathcal{P}_2\}, \mathcal{P}_1(h) \neq P_2(h)\}.$$

Under the proposed estimation approach we re-estimate the same Equation 1 that we used in the traditional approach, but only using customers in the *Different Recommendations Group (RBP)*. The estimated coefficient $\hat{\gamma}$ provides a conditional treatment effect for the group of customers where the two policies recommend different actions. This is the incremental performance of Policy 1 compared to Policy 2 among these customers.

To estimate an overall outcome, $\hat{V}(\mathcal{P}_1 - \mathcal{P}_2)$, we use a weighted average of the estimated differences in the two groups of customers, where the notation $\hat{V}(\mathcal{P}_1 - \mathcal{P}_2)$ denotes an estimator of $V(\mathcal{P}_1 - P_2)$. For the group of customers where the two policies recommend the same actions, we know that the difference in performance is exactly zero, with zero variance. We can therefore estimate $V(\mathcal{P}_1 - P_2)$ by re-weighting $\hat{\gamma}$ to adjust for the relative size of the two groups:[6]

$$\hat{V}(\mathcal{P}_1 - \mathcal{P}_2) = \frac{|h: \mathcal{P}_1(h) \neq \mathcal{P}_2(h)|}{H} \cdot \hat{\gamma}.$$

Weighting ensures that when we estimate the difference of the two policies, we do not ignore the group of customers for which the two policies recommend the same marketing action. We know that the difference in performance in this group is zero, and so not taking this group into account would result in positive bias in the absolute magnitude of the difference.

[6] This weight includes all of the customers, including customers in other experimental conditions (if any). Alternatively, we can calculate the weight when restricting attention to customers in the experimental conditions associated with the two policies.

We also obtain the standard error of the performance difference $\hat{V}(\mathcal{P}_1 - \mathcal{P}_2)$ by reweighting the standard error of the estimate of γ:

$$s.e.\left(\hat{V}(\mathcal{P}_1 - \mathcal{P}_2)\right) = \frac{|h: \mathcal{P}_1(h) \neq \mathcal{P}_2(h)|}{H} \cdot \ s.e.(\hat{\gamma})$$

Notice that when using RBP data, the only difference between the proposed and traditional approaches is that for the customers for whom $\mathcal{P}_1$ and $\mathcal{P}_2$ recommend the same marketing action, the proposed approaches fixes the performance difference at zero, with zero variance, while the traditional approach relies upon the observed performance differences. Because any observed performance differences reflect random noise, the proposed approach yields a more efficient estimate.

Extending the proposed estimation approach to RBA data is straightforward. We discuss this extension next.

Proposed Estimation Using RBA Data

We begin by constructing a sample of observations associated with each policy. The $Policy\ 1\ Dataset_{RBA}$ includes the RBA customers that were randomly assigned to receive the treatment recommended by $\mathcal{P}_1$:

$$Policy\ 1\ Dataset_{RBA} \equiv \{h: \mathcal{P}_1(h) = W_h\}$$

We define the $Policy\ 2\ Dataset_{RBA}$ similarly:

$$Policy\ 2\ Dataset_{RBA} \equiv \{h: \mathcal{P}_2(h) = W_h\}$$

We then use these two datasets to divide customers into two groups:

> **Same Recommendations Group (RBA)**: the set of customers in the union of the $Policy\ 1\ Dataset_{RBA}$ and the $Policy\ 2\ Dataset_{RBA}$ for whom both policies recommend the same action:
>
> $$\{h: \mathcal{P}_1(h) = \mathcal{P}_2(h) = W_h\}.$$
>
> **Different Recommendations Group (RBA)**: the set of customers in the union of the $Policy\ 1\ Dataset_{RBA}$ and the $Policy\ 2\ Dataset_{RBA}$ for whom the two policies recommend different actions:
>
> $$\{h: W_h = \mathcal{P}_1(h) \neq P_2(h)\} \cup \{h: W_h = P_2(h) \neq \mathcal{P}_1(h)\}.$$

Notice that when constructing these two groups, we restrict attention to customers that received the actions recommended by either $\mathcal{P}_1$ or $\mathcal{P}_2$ (customers in the union of the $Policy\ 1\ Dataset_{RBA}$ and the $Policy\ 2\ Dataset_{RBA}$). This restriction is important because in an RBA experimental design, there are some customers who receive marketing actions that are not recommended for them by either policy.

Having identified these two groups of customers, we then follow the same proposed approach that we use for the RBP data. In particular, we estimate Equation 1 using only the customers in the *Different Recommendations Group (RBA)*. In this setting, the *Policy 1* binary indicator identifies the customers in the $Policy\ 1\ Dataset_{RBA}$.[7] The estimate of γ provides a conditional treatment effect for the group of customers where the two policies recommend different actions. To obtain an average treatment effect, we again re-weight the estimate of γ to adjust for the number of observations in each group:

$$\hat{V}(\mathcal{P}_1 - \mathcal{P}_2) = \frac{|h: \mathcal{P}_1(h) \neq \mathcal{P}_2(h)|}{H} \cdot \hat{\gamma}$$

In this equation, the weighting factor can be calculated using the recommended actions for all of the customers in the RBA experiment.[8] The standard error of the performance difference $\hat{V}(\mathcal{P}_1 - \mathcal{P}_2)$ is the weighted standard error of the estimate $\hat{\gamma}$:

$$s.e.\left(\hat{V}(\mathcal{P}_1 - \mathcal{P}_2)\right) = \frac{|h: \mathcal{P}_1(h) \neq \mathcal{P}_2(h)|}{H} \cdot \ s.e.(\hat{\gamma})$$

To summarize, the difference in the proposed estimation approach when using RBA versus RBP data is the construction of the estimation sample. In an RBP dataset we restrict attention to the customers that are in the two experimental conditions associated with policies $\mathcal{P}_1$ and $\mathcal{P}_2$, whereas in an RBA dataset we restrict attention to customers that received the marketing action recommended by one of the two policies. In both cases, we then identify customers for which $\mathcal{P}_1$ and $\mathcal{P}_2$ recommend different marketing actions.

Additional Comments

Before presenting an application of the proposed approach we have two additional comments. First, because we are using OLS to compare the performance of $\mathcal{P}_1$ and $\mathcal{P}_2$, we recommend using standard regression techniques to improve the estimates of the standard errors. For example, if we believe the errors are correlated across observations we can cluster the standard errors. We can also use Eicker-Huber-White standard errors (Eicker, 1967; Huber, 1967; White, 1980) to correct for heteroscedasticity.

Second, we have shown how the proposed estimation approach can be used with either RBA or RBP data. However, we have only described the use of the traditional estimation method with RBP data. It is not clear how to use the traditional approach with RBA data in the framework of OLS estimation. Recall that the traditional approach estimates Equation 1 when including customers for which $\mathcal{P}_1$ and $\mathcal{P}_2$ recommend the same marketing actions. In an RBP design, some of these customers (for whom the two policies

[7] When using data from an RBP experiment, the *Policy 1* indicator identifies customers in the treatment condition associated with $\mathcal{P}_1$.
[8] This includes customers who were randomly assigned to receive marketing actions that are different than the actions recommended by $\mathcal{P}_1$ and $\mathcal{P}_2$.

recommend the same action) are randomly assigned to the treatment associated with $\mathcal{P}_1$, and the others are randomly assigned to the treatment associated with $\mathcal{P}_2$. In contrast, in an RBA design these customers are not assigned to a specific policy; when two policies recommend the same action in an RBA design, the same customers are used to evaluate each policy. This makes the construction of the *Policy 1* indicator ambiguous when using RBA data under the traditional approach.[9]

In the next section we illustrate the proposed approach using data from an actual field experiment. We will show empirically that the efficiency improvements can be large. Our proposed estimation method reduces the standard error by more than 25%.

5. Application to An Actual Field Experiment

Simester, Timoshenko, and Zoumpoulis (2019) (hereafter "STZ") investigate how a retailer should target prospective customers. They consider three marketing actions; a discount, a free trial and a no-mail control. We label these marketing actions: *Discount*, *Free Trial* and *Control*. STZ compare seven optimized targeting policies, which each assign one of the three marketing actions to every customer.

The STZ study included approximately four million prospective households grouped into carrier routes (approximately 400 households per carrier route). There are thirteen covariates describing the characteristics of each carrier route. The covariates vary at the carrier route level (there is no observed variation within a carrier route). The seven targeting policies were all trained using these covariates, and so for each targeting policy the recommended marketing actions vary across carrier routes but do not vary across households within the same carrier route. We provide more information on the covariates in Appendix B.

The carrier routes were assigned into ten experimental conditions. The randomization was conducted at the carrier route level, so that all of the households in the same carrier route received the same treatment. The ten experimental conditions include seven treatments in an RBP (randomized-by-policy) experimental design. In these seven treatments, carrier routes were randomly assigned to one of the seven targeting policies; all of the households in a carrier route assigned to a targeting policy received the marketing action recommended for that carrier route by that targeting policy. The other three randomly assigned treatments use the RBA (randomized-by-action) experimental design. The three treatments include one condition in which all of the carrier routes received the *Discount*, another condition in which they all received the *Free Trial*, and a third condition in which they did not receive any promotion (the *Control*). In our analysis we will treat the unit of observation as a carrier route (aggregating outcomes

[9] One option would be to randomly assign the *Policy 1* indicator for this group of customers. However, this random assignment would introduce an additional source of noise, affecting both the point estimate and the standard errors. Another option would be to duplicate the observations for this group of customers, although this would require careful treatment of the standard errors.

across households within each carrier route). In Table 1 we summarize the design of the STZ study using the notation introduced in Sections 3 and 4. For ease of exposition, in the RBP data we restrict attention to two of the seven experimental conditions, which we label "Policy 1" and "Policy 2".

Table 1. Mapping Notation to the STZ Study

Notation	Description	RBP Data (Policy 1 and 2)	RBA Data
$\mathcal{A}$	Set of marketing actions	*Discount* *Free Trial* *Control*	*Discount* *Free Trial* *Control*
h	Unit of observation	A carrier route	A carrier route
$Y_h(a)$	Outcome measure	Profit	Profit
$\boldsymbol{x}_h$	Vector of covariates for carrier route h	13 covariates	13 covariates
L	Number of experimental conditions	2	3
H	Total number of carrier routes	2,122	3,091
	Number of carrier routes assigned to each experimental condition	Policy 1 = 1,046 Policy 2 = 1,076	*Discount* = 1,003 *Free Trial* = 1,026 *Control* = 1,062

We first illustrate how to evaluate a single policy using RBA data. In Table 2 we group the carrier routes assigned to the three RBA conditions using the recommended actions for Policy 1. Across all three conditions, there are a total of 2,269 carrier routes (764 + 741 +764) for which Policy 1 recommended sending the *Discount*. Within this group of 2,269 carrier routes, there is a subgroup of 764 that actually received the *Discount.* We can use these 764 carrier routes to evaluate the outcome for this group. The subgroups used to evaluate the other two marketing actions are highlighted by shading in Table 2. Pooling these subsamples yields a total of 1,061 carrier routes that can be used to evaluate Policy 1.

Table 2. Evaluating a Single Policy

Policy 1 Recommendation	Discount Condition	Free Trial Condition	Control Condition
Discount	764	741	764
Free Trial	6	5	6
Control	233	280	292
Total	**1,003**	**1,026**	**1,062**

The table reports the number of carrier routes in the three RBA experimental conditions in the STZ study. The observations are grouped by the actions recommended by Policy 1. The shading identifies the observations used to evaluate the performance of Policy 1.

We can also illustrate how to compare the performance of Policy 1 and Policy 2. In Table 3 we group the RBA and RBP observations according to whether the two policies recommend the same or different actions. In the RBA data, there are 492 carrier routes for which the two policies recommend the same actions. For these carrier routes the true difference in the performance of the two policies is zero.[10] Our proposed approach recognizes this, and so just focuses on the performance difference in the carrier routes where the two policies recommend different actions. We then weight the performance difference for these carrier routes to adjust for the relative size of the two groups.

Table 3. Comparing Two Policies

			Recommend Same Actions	Recommend Different Actions	All Carrier Routes
RBA Design	Policy 1	Number of Carrier Routes	492	569	1,061
		Average Profit	$12.405	$11.925	$12.148
	Policy 2	Number of Carrier Routes	492	532	1,024
		Average Profit	$12.405	$10.754	$11.547
RBP Design	Policy 1	Number of Carrier Routes	502	544	1,046
		Average Profit	$11.400	$13.887	$12.694
	Policy 2	Number of Carrier Routes	512	564	1,076
		Average Profit	$12.214	$10.409	$11.268

The table reports the average profits from the three RBA experimental conditions, and the two RBP experimental conditions associated with Policy 1 and Policy 2, in the STZ study. To preserve confidentiality, the profits are multiplied by a common random number. The observations are grouped according to whether Policy 1 and Policy 2 recommended the same or different actions.

If instead we used the RBA data to calculate the overall average profit for each policy without separating the two groups, we would compare the outcome for the 1,061 carrier routes used to evaluate Policy 1 with the 1,024 carrier routes used to evaluate Policy 2. Notice that random variation means that Policy 1 and 2 have slightly different numbers of observations when they recommend different actions: 569 vs. 532 carrier routes (in the RBA data). As a result, when comparing the overall average profit ($12.148 versus $11.547), the profits for the 492 carrier routes (for which the two policies recommend the same actions) will not perfectly cancel out. This introduces random error, which our proposed approach removes.

[10] If we were only interested in the relative performance of these two policies (and no other policies), and were not interested in the absolute performance of either policy, then we could omit from the study these 492 carrier routes. These carrier routes provide no information about the relative performance of these two policies. The costs associated with treating these carrier routes could either be saved, or re-allocated to carrier routes for which the policies recommend different actions.

Distinguishing between the two groups of carrier routes is even more important when using RBP data. The traditional estimation approach compares the overall average response in the 1,046 carrier routes assigned to Policy 1, with the overall average response in the 1,076 carrier routes associated with Policy 2. This fails to recognize that for many of the carrier routes there is no difference in the performance of the two policies. In particular, there are 502 carrier routes assigned to Policy 1 and 512 carrier routes assigned to Policy 2 for which there is no difference between the two policies. However, random noise suggests there is a difference between the two policies among these carrier routes ($11.400 vs. $12.214). The proposed estimation approach would remove this noise, by recognizing that the true difference is zero and focusing instead on the group of carrier routes for which the policies make different recommendations. For completeness, we showcase this calculation in detail in Appendix C. In the remainder of this section we measure the magnitude of the efficiency improvements.

Efficiency Improvements

We contrast the findings under the proposed estimation approach versus the traditional estimation approach. Because the STZ study implemented both an RBP and an RBA design, we also compare the results from the two experimental designs. We start by using the RBP data to calculate the traditional and the proposed estimates described in Section 4. Comparing the results under these two approaches reveals the efficiency improvement from recognizing that there is no difference in the performance of the two policies when they recommend the same action. We then demonstrate the proposed estimation approach using the RBA data to evaluate whether our proposed estimation approach yields similar findings when using either RBP or RBA data. We report the findings in Table 4, where we report the analytical standard errors from OLS adjusted for heteroscedasticity using the Eicker-Huber-White adjustment.

Table 4. Comparing Experimental Designs and Estimation Approaches

Experimental Design	Traditional Estimation (No Grouping)	Proposed Estimation (Grouping)
RBP	$1.002 ($0.805)	$1.506 ($0.600)
RBA		$0.988 ($0.544)

Note: The table reports the estimated average difference in the performance of Policy 1 and Policy 2 using data from both experimental designs. Standard errors are in parentheses. The unit of analysis is a carrier route. The OLS standard errors are adjusted for heteroscedasticity using the Eicker-Huber-White adjustment. To preserve confidentiality, the profits are multiplied by a common random number.

There are two findings of interest. First, the proposed estimation approach reduces the standard errors by over 25% (compared to the traditional approach). This efficiency improvement has substantive

importance; the difference in the performance of the two policies becomes statistically significant, whereas the difference is not significant when using the traditional estimation approach (no grouping). Recall that when grouping, the difference in the performance of the policies is set to zero when two policies recommend the same action. This is the source of the efficiency gains.[11]

Second, the standard errors under the proposed estimation approach are very similar for the two experimental designs. However, to compare all seven policies, the RBP design requires seven experimental conditions, while the RBA design requires just three. If we just had access to the RBP data for Policy 1 and Policy 2, we would not be able to evaluate the performance of any of the other five targeting policies using the RBP data. Using the RBA data we can evaluate any of the seven targeting policies, or any other policy (subject to footnote 1).

Summary

In this section we used data from the STZ study to illustrate the benefits from the proposed approach. The findings reveal that the RBA experimental design generates qualitatively similar conclusions about the performance of the seven tested policies as a traditional RBP design. However, it accomplishes this goal using just three experimental conditions instead of seven.The findings also confirm that the proposed estimation approach yields a more precise comparison of two policies. The standard error of the estimates of the performance difference is reduced by over 25%.

Implementing an experiment to compare targeting policies yields data that can be used for more than just validation; the data can also be used to train new policies. In the next section we investigate whether the cornerstone of our estimation approach, grouping customers using the recommended marketing actions, can also help to improve the training of new policies.

6. Training New Targeting Policies

For the firm that provided the data in the STZ study, there are important advantages of using the data in the three RBA treatments to train a new policy. The data provides an additional source of information over the older information used to train the seven candidate policies.[12] Moreover, STZ document that

[11] In our application, the variance reduction from adding covariates is a lot less than the variance reduction from setting the difference to zero for customers for whom the policies recommend the same actions. This is consistent with the findings of Johnson, Lewis, Reiley (2017).

[12] The STZ study actually included two experiments conducted approximately six months apart. The first experiment provided data to train the seven targeting policies, and the second experiment was used to compare the performance of the seven policies. We used data from the second experiment in Section 5 to illustrate the efficiency advantages of our proposed estimation approach.

there is evidence of non-stationarity, so that the data from this experiment is more representative of current market conditions (than the older training data).

One option would be to focus solely on data from the three RBA treatments when training a new targeting policy. However, training the new policy solely using this data may mean losing some of the older information that was used to create the seven policies. Even though there is evidence of non-stationarity, some of the information in the older data is likely to be valuable when training new policies. In this section we describe a way to preserve and incorporate this information when using the new data to train new policies. This is particularly important if the existing policies were trained using intuition or datasets that are no longer available, and so there is a risk that the information in the existing policies may be lost.

For ease of exposition, we label the RBA data from the STZ study used in Section 5 as the "new" training data. For this data we have thirteen covariates to use to target which customers should receive the *Discount* and *Free Trial* promotions (or *Control*). For each customer in the new training data we also log which promotion (*Discount*, *Free Trial*, or *Control*) Policy 1 and Policy 2 recommend sending. We can now train new targeting policies using just the thirteen targeting variables, or using these thirteen targeting variables *plus* the logged recommended actions.

As a preliminary step, we first construct a "standard" new policy as a benchmark using the RBA experimental conditions in the new training data. We label this new policy "Standard Lasso" and compare this benchmark to Policy 1 and Policy 2 (the existing policies used in Section 5). In particular, we implement the following procedure:

Step 1 Construct Data: randomly divide the new RBA data into calibration (70%) and validation (30%) subsamples.

Step 2 Train New Model: Use the calibration subsample to train a new policy ("Standard Lasso"):

i. Use Lasso to separately estimate three predictive models m_a corresponding to each of the marketing actions a:

$$Y_h(a) = m_a(\boldsymbol{x}_h),$$

where $\boldsymbol{x}_h$ represents the thirteen covariates and $a \in \{Discount, Free\ Trial, Control\}$.

ii. Use the three models to predict the profit for each marketing action for each household in the validation subsample:

$$\{\widehat{Y_h}(Discount), \widehat{Y_h}(Free\ Trial), \widehat{Y_h}(Control)\}$$

iii. For every observation h in the validation subsample, the targeting policy assigns the marketing action a that yields the highest predicted profit.

Step 3 Evaluate Benchmarks: Use the validation subsample and our proposed estimation approach (focusing on customers for whom the two policies recommend different actions) to compare:

a. Policy 1 with Standard Lasso
b. Policy 2 with Standard Lasso

Notice that Standard Lasso is trained solely using the new RBA data. We repeat the procedure 1,000 times using different random draws for calibration and validation in Step 1. Table 5 reports the average difference in profit between the two existing policies and Standard Lasso (to preserve confidentiality, we multiply the profits by a common random number). As expected, Standard Lasso outperforms both of the existing policies. This is not surprising; Standard Lasso is calibrated using the new data, which matches the (new) evaluation data. In contrast, the existing policies were trained using old data.

Table 5. Average Profits From the Existing and New Policies Compared to "Standard Lasso"

		Average Profit	Standard Error
Existing Policies	Policy 1	-$0.257	$0.021
	Policy 2	-$0.956	$0.030
New Policies	Lasso with Policy 1	$0.015	$0.011
	Lasso with Policy 2	$0.003	$0.006
	Lasso with Both Policies	$0.028	$0.012

The first two rows of the table compare Policy 1 and Policy 2 to Standard Lasso. The last three rows compare new policies that incorporate information from the existing policies to Standard Lasso. Standard Lasso is the policy trained using the new STZ experimental data. Negative (positive) values indicate that Standard Lasso is more (less) profitable than the other policies. The profit differences are averaged across 1,000 Monte-Carlo cross-validation iterations. In each iteration, we split data into calibration and validation subsamples (70%:30%). We train the targeting methods on the calibration data and evaluate performance on the validation data. The profits are multiplied by a common random number.

To investigate whether information from the existing policies can improve the new policy, we define four indicator variables. *Policy1_Discount* is an indicator variable that equals one if Policy 1 recommends the *Discount* (and equals zero otherwise). We similarly define *Policy1_ FreeTrial*, *Policy2_Discount*, and *Policy2_FreeTrial*. We then use the same random draws of the calibration sub-sample to train three new targeting policies by adding these indicator variables to the training dataset. To train these new policies we use Lasso to estimate $Y_h(a) = m_a(\boldsymbol{x}_h, \boldsymbol{b})$, where $\boldsymbol{x}_h$ represents the thirteen covariates and $\boldsymbol{b}$ is the

vector of binary indicators identifying the recommended actions from the existing policies.[13] We then evaluate these new policies using the validation subsample. The findings are also reported in Table 5.

Adding the binary indicators from both Policy 1 and 2 to the thirteen covariates yields a significantly more profitable targeting model.[14] We conclude that logging the recommended actions from existing policies provides a simple way to improve new targeting policies by incorporating information from existing policies. The proposed approach does not require merging old and new datasets and can use any existing targeting policy, including policies developed using intuition or data that is no longer available.

While the proposed experimental design offers important benefits, it also has limitations. We discuss these limitations next.

7. Limitations of Randomizing by Action

One potential disadvantage of randomly assigning customers to marketing actions is cost. It is sometimes obvious that a marketing action is optimal for only a small portion of the population, and so randomly assigning customers to receive this action may lead to an opportunity cost. The data from the STZ study highlights this cost. In the STZ study, the *Discount* is more profitable than the other two marketing actions. The seven optimized policies recognize the profitability of the *Discount*, and so they recommend sending a *Discount* to most households. As a result, the average profit across the approximately 2.8 million households assigned to the seven RBP experimental conditions is significantly higher than the average profit earned from the approximately 1.2 million households randomly assigned to the three marketing actions. Randomly assigning these 1.2 million customers to marketing actions resulted in an opportunity cost to the firm of over $100,000. This cost could be reduced by underweighting the *Free Trial* and *Control* when randomly assigning customers.

In a related point, it may be unethical or unacceptable to randomly assign some customers to some marketing actions. This limitation is easily addressed by designing the randomization procedures to prevent experimental conditions that are unacceptable or unethical. Although this may prevent evaluation of every possible policy, it allows evaluation of any policy that is acceptable and ethical.

We also recognize that the feasibility of the proposed program depends upon the size of the action space. If the action space is too large then it may not be feasible to implement the proposed design, and/or we

[13] We label these new policies *Lasso with Policy 1*, *Lasso with Policy 2*, and *Lasso with Both Policies*. For *Lasso with Policy 1*, $\boldsymbol{b}$ includes *Policy1_Discount* and *Policy1_FreeTrial*. For *Lasso with Policy 2*, $\boldsymbol{b}$ includes *Policy2_Discount* and *Policy2_FreeTrial*. *Lasso with Both Policies* incorporates all four indicator variables.
[14] We confirmed the robustness of this finding by varying which machine learning method we used to train the new policies, and which of the seven existing policies we considered. Using indicator variables to incorporate information from the existing policies consistently improved the performance of the new policy.

may have too few observations to accurately estimate the conditional treatment effects. This limitation is particularly relevant for dynamic policies that involve a sequence of decisions. Each unique sequence of actions represents a single "marketing action". For example, Simester et al. (2006) tested their dynamic catalog targeting policy using a sequence of twelve catalog mailing opportunities. With a mail or no mail decision on each mailing opportunity, this yielded an action space with 4,096 possible marketing actions. Potential solutions include supplementing the experimental data with historical data, particularly where the same marketing actions were also implemented in the past. Interpolation may also allow the removal of some intermediate actions from the experimental design.

8. Conclusions

We have presented an approach to designing and analyzing targeting experiments that offers three important advantages. First, the proposed experimental design allows evaluation (and comparison) of any policies, including policies designed after the experiment is implemented. Second, our approach yields more efficient estimates of the difference in the performance of the policies. Third, the proposed approach offers opportunities to improve targeting policies. We illustrated these benefits using data from an actual field experiment. The findings confirm that the benefits can be substantial.

References

Dubé, J.-P. and S. Misra (2017), "Scalable Price Targeting," working paper, University of Chicago.

Dudík, M., J. Langford, and Lihong Li (2011), "Doubly Robust Policy Evaluation and Learning," *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 1097-1104.

Eicker, Friedhelm (1967), "Limit Theorems for Regression with Unequal and Dependent Errors," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 59-82.

Hitsch, G. and S. Misra (2018), "Heterogeneous Treatment Effects and Optimal Targeting Policy," working paper, University of Chicago.

Horvitz, D. G.; Thompson, D. J. (1952) "A Generalization of Sampling Without Replacement From a Finite Universe", *Journal of the American Statistical Association*, 47, 663–685

Huber, Peter J. (1967). "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 221–233.

Johnson, G. A., R. A. Lewis, and E. I. Nubbemeyer (2017), "Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness," *Journal of Marketing Research*, 54(6), 867-884.

Johnson, G., R. A. Lewis, and D. H. Reiley (2017), "When Less Is More: Data and Power in Advertising Experiments," *Marketing Science*, 36, 1, 43–53.

Langford, J., A. Strehl, and J. Wortman (2008), "Exploration Scavenging," *Proceedings of the 25th International conference on Machine Learning*, 528-535.

Mantrala, Murali K., P. B. Seetharaman, Rajeeve Kaul, and Srinath Gopalakrishna, and Antonie Stam (2006), "Optimal Pricing Strategies for an Automotive Aftermarket Retailer," *Journal of Marketing Research*, Vol. XLIII, November, 588-604.

Ostrovsky, Michael and Michael Schwarz (2011), "Reserve Prices in Internet Advertising Auctions: A Field Experiment," in *Proceedings of the 12th ACM conference on Electronic Commerce* (EC '11). ACM, New York, NY, USA, 59-60.

Rafieian, O. and H. Yoganarasimhan (2018), "Targeting and Privacy in Mobile Advertising," working paper, University of Washington.

Simester, D., A. Timoshenko and S. I. Zoumpoulis (2019), "Targeting Prospective Customers: Robustness of Machine Learning Methods to Typical Data Challenges," *Management Science*, forthcoming.

Skiera, Bernd, and Nadia Abou Nabout (2013), "PROSAD: A Bidding Decision Support System for Profit Optimizing Search Engine Advertising," *Marketing Science*, 32(2), 213-220.

Strehl, A. L., J. Langford, L. Li, S. M. Kakade (2010), "Learning from Logged Implicit Exploration Data," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, Volume 2, 2217-2225.

Sutton, Richard S. and Andrew G. Barto (1998), *Introduction to Reinforcement Learning*, 1st edition, MIT Press, Cambridge, MA, USA.

White, Halbert (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*. **48** (4): 817–838.

Appendix A. Comparison of Policies Using Direct Estimation

As an alternative to the OLS-based estimation we present in Section 4, we also propose a direct estimation method for comparing policies. Both in the OLS-based estimation, and in the direct estimation approach, when comparing two targeting policies $\mathcal{P}_1$ and $\mathcal{P}_2$, we recognize that the true difference of the policies is zero (with zero variance) in the group of customers where the two policies recommend the same actions. In particular, we propose evaluating the difference using only the group of customers where the policies recommend different actions.

When using RBP data, we propose the estimator:

$$\hat{V}(\mathcal{P}_1 - \mathcal{P}_2) = \delta_{\mathcal{P}_1 \neq \mathcal{P}_2} \left[\frac{1}{|h: W_h = \mathcal{P}_1, \mathcal{P}_1(h) \neq \mathcal{P}_2(h)|} \sum_{h: W_h = \mathcal{P}_1, \mathcal{P}_1(h) \neq \mathcal{P}_2(h)} Y_h^{obs} - \frac{1}{|h: W_h = \mathcal{P}_2, \mathcal{P}_1(h) \neq \mathcal{P}_2(h)|} \sum_{h: W_h = \mathcal{P}_2, \mathcal{P}_1(h) \neq \mathcal{P}_2(h)} Y_h^{obs} \right],$$

where the weight $\delta_{\mathcal{P}_1 \neq \mathcal{P}_2}$ is given by:

$$\delta_{\mathcal{P}_1 \neq \mathcal{P}_2} = \frac{|h: \mathcal{P}_1(h) \neq \mathcal{P}_2(h)|}{H}$$

When using RBA data, we propose the estimator:

$$\hat{V}(\mathcal{P}_1 - \mathcal{P}_2) = \delta_{\mathcal{P}_1 \neq \mathcal{P}_2} \left[\frac{1}{|h: W_h = \mathcal{P}_1(h) \neq \mathcal{P}_2(h)|} \sum_{h: W_h = \mathcal{P}_1(h) \neq \mathcal{P}_2(h)} Y_h^{obs} - \frac{1}{|h: W_h = \mathcal{P}_2(h) \neq \mathcal{P}_1(h)|} \sum_{h: W_h = \mathcal{P}_2(h) \neq \mathcal{P}_1(h)} Y_h^{obs} \right].$$

The weighting ensures that when we estimate the difference of the two policies, we do not ignore the group of customers where the two policies recommend the same actions. We know that the difference in performance in this group is zero, and so not taking this group into account would result in positive bias in the absolute magnitude of the difference.

Appendix B. Additional Information on the Thirteen Covariates

Definitions of Targeting Variables

Variable	Definition
Age	Age of head of household
Home Value	Estimated home value
Income	Estimated household income
Single Family	A binary flag indicating whether the home is a single family home
Multi-Family	A binary flag indicating whether the home is a multi-family home
Distance	Distance to nearest store for this retailer
Comp. Distance	Distance to nearest competitors' store
Penetration Rate	% of households in zip code that are members
3yr Response	Average response rate to mailings to this zip code over the last 3 years
F Flag	Binary flag indicating whether the retailer considers the zip code "far" from its closest store
M Flag	Binary flag indicating whether the retailer considers the zip code a "medium" distance from its closest store
Past Paids	The proportion of households in the zip code that were previously paid members
Trialists	The proportion of households in the zip code that have been identified as households who repeatedly sign up for trial memberships

The demographic variables were purchased by the retailer from a third-party commercial data supplier. The remaining variables were constructed by the retailer using the retailer's own data.

The strongest indicator that a carrier route will yield large profits is a high previous response rate (*3yr Response*). Other significant factors indicating larger expected profits include: a short distance to the nearest own store (*Distance*), a long distance to the competitors' store (*Competitive Distance*), a concentration of single family housing (*Single Family*), a low average age (*Age*), and a high proportion of households that were previously paid members (*Past Paids*).

Additional details can be found in Simester, Timoshenko and Zoumpoulis (2019).

Appendix C. STZ Study: Comparison of Policies Using RBP Data

Under the traditional approach, the difference between Policy 1 and Policy 2 can be estimated as

$$\begin{aligned}\hat{V}(\mathcal{P}_1) - \hat{V}(\mathcal{P}_2) &= \left(\frac{502}{1{,}046} \cdot \$11.400 + \frac{544}{1{,}046} \cdot \$13.887\right) - \left(\frac{512}{1{,}076} \cdot \$12.214 + \frac{564}{1{,}076} \cdot \$10.409\right) \\ &= \underbrace{\left(\frac{502}{1{,}046} \cdot \$11.400 - \frac{512}{1{,}076} \cdot \$12.214\right)}_{\neq 0} + \left(\frac{544}{1{,}046} \cdot \$13.887 - \frac{564}{1{,}076} \cdot \$10.409\right).\end{aligned}$$

Under the proposed approach, and using the direct estimation method, we can write

$$\hat{V}(\mathcal{P}_1 - \mathcal{P}_2) = \underbrace{\left(1 - \delta_{\mathcal{P}_1 \neq \mathcal{P}_2}\right) \cdot 0}_{=0} + \delta_{\mathcal{P}_1 \neq \mathcal{P}_2} \cdot (\$13.887 - \$10.409),$$

where $\delta_{\mathcal{P}_1 \neq \mathcal{P}_2} = \frac{544+564}{1{,}046+1{,}076}$ is the share of carrier routes for which the two policies recommend different actions.