

Science of Experimentation at Amazon

Joe Coopriider and Shima Nassiri

1 Introduction

In order to test new pricing policies and improve prices at Amazon, we created an online pricing experimentation service that helps teams measure the causal impact of their changes on strategies/policies that affect prices seen by customers on Amazon. Since we do not price discriminate (i.e., we do not show different prices to different customers at the same time), we must run product-randomized experiments. Our service supports online A/B tests through statistical hypothesis testing to measure incremental effects (other terms loosely describing such experiments include randomized control trails, z-tests, etc.) of experiments run in real time on the [Amazon.com](https://www.amazon.com) website.

In this paper, we describe 1) what we do as scientists is to improve the functionality of our pricing service, 2) how we help lab owners design their experiments and understand the analysis results, 3) how we increase precision through improved experimental design (i.e., crossovers), and better estimators that control for demand trends and differences between treatment groups, and 4) ways to reduce bias by improving randomization to prevent spillovers.

2 Overview

In order to run experiments to measure the impact of prices on customers, we randomize products into treatment group(s) and a control group, where the treatment group is priced by the new pricing policy and the control group is priced by the existing pricing policy. The purpose of pricing experimentations is to estimate the average treatment effect (ATE) of a pricing policy to determine whether the policy should be launched is generally not to measure price elasticity.

The pricing experiments can be categorized into two types. The first type is time-bound experiments where products will be treated throughout the entire experimental period. Consider you want to test a change in a ML algorithm that sets the price of a group of products. For experiments like this, we have a baseline period where no products are treated (i.e. they are priced using the existing ML algorithm). At the start of the experiments, products assigned to the treatment group receive would be priced using the updated ML algorithm while the control products remain priced using the existing ML algorithm. At the end of the experiment, products in the treatment group are compared to those in the control group to measure the ATE. An illustration is shown in Figure 1 below, where the red cells are when a product is being treated.

Figure 1: Time-Bound Experimental Design

| | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | Day 6 | Day 7 | Day 8 | Day 9 | Day 10 | Day 11 | Day 12 | Day 13 | Day 14 |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|--------|--------|--------|--------|
| Product A | | | | | | | | | | | | | | |
| Product B | | | | | | | | | | | | | | |
| Product C | | | | | | | | | | | | | | |
| Product D | | | | | | | | | | | | | | |
| Product E | | | | | | | | | | | | | | |
| Product F | | | | | | | | | | | | | | |
| Product G | | | | | | | | | | | | | | |
| Product H | | | | | | | | | | | | | | |
| Product I | | | | | | | | | | | | | | |
| Product J | | | | | | | | | | | | | | |

33 Time-bound experiments are not always the best design for pricing experiments at Amazon. Our prices
 34 fluctuate based on a variety of different factors that can change over time (e.g., costs, promotions, prices
 35 at other stores etc). Some of these can change during our experiment regardless of what policy we are
 36 testing. Changes in factors that determine our prices during the experiment can change the experiment
 37 population. We use trigger-based experiments for these cases. A trigger-based experiment is when a
 38 product is only in the experimental analysis after a “trigger” is met. This means that only a subset of the
 39 original experiment population is analyzed. Once a product is triggered, it enters the experiment
 40 regardless of the treatment group it belongs to. If the product is in the treated group, the new policy will
 41 be applied to it after being triggered, while products in the control group continue with the existing
 42 policy. When products get triggered, we consider them as triggered until the end of the experiment.

43 Below is an example of a trigger-based experiment. Red cells are treated experimental periods and green
 44 cells are control experimental periods. Suppose the trigger is a product being put on promotion at
 45 another store. Once a product is triggered (another store puts it on promotion), it enters the experiment
 46 and is assigned to either treatment or control. In this example, products A and E are on promotion at
 47 another store on day 8 (the first day of the experiment). Products B, F, G, I and J are put on promotion at
 48 another store during the experiment and are added to the analysis the day that they are put on promotion.
 49 Note that in this example, three products (C, D, and H) are never triggered and are left out of the
 50 analysis.

51 **Figure 2: Trigger-Based Experimental Design**

| | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 | Day 6 | Day 7 | Day 8 | Day 9 | Day 10 | Day 11 | Day 12 | Day 13 | Day 14 |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|--------|--------|--------|--------|
| Product A | | | | | | | | Red | Red | Red | Red | Red | Red | Red |
| Product B | | | | | | | | | | Green | Green | Green | Green | Green |
| Product C | | | | | | | | | | | | | | |
| Product D | | | | | | | | | | | | | | |
| Product E | | | | | | | | Green | Green | Green | Green | Green | Green | Green |
| Product F | | | | | | | | | | | | | Red | Red |
| Product G | | | | | | | | | | | Red | Red | Red | Red |
| Product H | | | | | | | | | | | | | | |
| Product I | | | | | | | | | | | Green | Green | Green | Green |
| Product J | | | | | | | | | Red | Red | Red | Red | Red | Red |

52

53 3 Improving Precision

54 Because of factors that we cannot detail in this manuscript, such as promotions, advertisements,
 55 influencer recommendations, or supply chain problems, product demand can have high variation. The
 56 noisy data environment in pricing experiment often leads to noisy ATE estimates in the product level
 57 experiment results. Noisy ATE estimates create confusion for the partner teams working with pricing
 58 experimentation service. To improve our precision, we have begun using a better experimental design
 59 called crossovers and developed a more precise estimator called the Heterogeneous Panel Treatment
 60 Effect (HPTE).

61 **3.1 Experimental Design**

62 Switchbacks are a common tool to improve precision and power in experiments. For this paper, we
 63 define switchbacks as when treatment varies across products and time during our experiment.
 64 Switchbacks generally occur when the treatment turns on or off multiple times for each product in the
 65 experiment. For the time-bound example above, products would switch between the new ML algorithm

66 price and the old ML algorithm price multiple times throughout the experiment. This is beneficial for
 67 many reasons: it exposes more products to treatment since each product can be treated during the
 68 experiment, it increases the variation of when the treatment is applied to each product since the start of
 69 the treatment is different for different products, and it provides a more effective counterfactual since
 70 each product has both treatment and control periods during the experiment.

71 In this setting, since the treatments start on different days for different products, it allows us to separate
 72 demand shifts across Amazon or among groups of products from the treatment more effectively. Further,
 73 we will have days within the experiment where each product is not treated, which will provide a more
 74 effective counterfactual than the control group or the treated group before the experiment began, the
 75 standard counterfactual periods in normal A/B tests. Below is an example of a switchback design called
 76 random days which randomly assigns each product-day to either treatment or control. Random days
 77 experiments can shrink standard errors by about 60%.

78 **Figure 3: Random-Days Experimental Design**

| Day: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|-----------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|
| Product A | | | | | | | | ■ | ■ | | | ■ | ■ | | ■ | | | ■ | | | ■ |
| Product B | | | | | | | | ■ | | | ■ | ■ | | | | ■ | ■ | ■ | | | ■ |
| Product C | | | | | | | | | ■ | ■ | | ■ | | | ■ | | | ■ | | | ■ |
| Product D | | | | | | | | | | ■ | ■ | | ■ | | | ■ | ■ | | | ■ | ■ |
| Product E | | | | | | | | ■ | | | ■ | | ■ | | ■ | | | ■ | | | ■ |
| Product F | | | | | | | | ■ | ■ | | ■ | | ■ | | ■ | | | ■ | | | ■ |
| Product G | | | | | | | | | ■ | ■ | | ■ | | ■ | | | ■ | | | ■ | ■ |
| Product H | | | | | | | | | | ■ | ■ | | ■ | | ■ | | | ■ | | | ■ |
| Product I | | | | | | | | ■ | | | ■ | | ■ | | ■ | | | ■ | | | ■ |
| Product J | | | | | | | | ■ | ■ | | | ■ | ■ | | | ■ | | | ■ | | ■ |

79

80 Random-days ATE estimates are only accurate if the prices on one day do not effect demand the
 81 following day. That is not the case in our environment. Lowering price one day, can lead to higher
 82 demand the next day. Higher demand one day can lead to increased traffic the following day through
 83 customer traffic mechanisms like search queries and recommended product widgets that may have past
 84 customer demand as an input. This is called the carry-over effect. Therefore, random-days ATE
 85 estimates can be biased as the treatment can affect the demand during control periods.

86 Under the crossover design, we split the experimental population into two groups: A and B. Group A is
 87 treated and Group B is control for the first half of the experiment. In the second half, Group B is treated
 88 and Group A is control. To minimize the bias from carry-over effect, we consider a blackout period at
 89 the beginning of the first and second half of the experiment.

90 Below is an example of crossover experimental design where week 7 is the start of the experiment and
 91 week 10 is the start of the second half of the experiment. Weeks 7 and 10 are blacked out because they
 92 are dropped from our analysis, but have the same treatment status as weeks 8 and 11 respectively.
 93 Consider we are comparing two ML algorithms. Group A would be priced with the new ML algorithm
 94 from week 7 to week and the old algorithm from week 10 to week 12. Group B would be priced with the
 95 old algorithm for weeks 7-9 and the new algorithm for weeks 10-12. Our analysis would not include
 96 weeks 7 and 10 because those effects could be biased by carryover effect from the prices in the previous
 97 week.

98

Figure 4: Crossover Experimental Design

| | Week 1 | Week 2 | Week 3 | Week 4 | Week 5 | Week 6 | Week 7 | Week 8 | Week 9 | Week 10 | Week 11 | Week 12 |
|-----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|---------|---------|
| Product A | | | | | | | | | | | | |
| Product B | | | | | | | | | | | | |
| Product C | | | | | | | | | | | | |
| Product D | | | | | | | | | | | | |
| Product E | | | | | | | | | | | | |
| Product F | | | | | | | | | | | | |
| Product G | | | | | | | | | | | | |
| Product H | | | | | | | | | | | | |
| Product I | | | | | | | | | | | | |
| Product J | | | | | | | | | | | | |

99

Crossover experiments shrink standard errors by about 40-50%. That is not as much as random-days but avoids potential carry-over effect. It is still effective, because it has most of the benefits of a random-days design as every product is exposed to both treatment and control during the experiment. This design can only be done for time-bound experiments and cannot be done for triggered-based experiments.

105

3.2 Heterogeneous Panel Treatment Effect (HPTE)

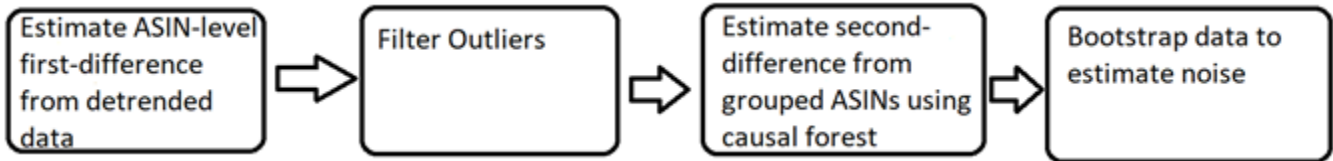
106

The goal of HPTE is to estimate the ATE of a new pricing policy compared to an existing policy at Amazon. The intuition behind HPTE follows the difference-in-difference (DID) structure: First, we use time-series data to identify the first difference of each product. Second, match similar products across the treatment and control group to take the difference of similar products' first difference (i.e., second difference).

111

The empirical steps of HPTE estimator are as follows: 1) Detrend the data using pre-experimental product-level trends; 2) Filter outlier products from detrended data; 3) Use causal forests to nonparametrically control for differences between treatment and control group; 4) Resample the data and estimate the ATE on the resampled data (bootstrapping) to estimate the distribution of possible ATE. A flowchart is shown below:

116



117

118

119

3.2.1 Methodology

120

3.2.1.1 Step 1: Estimate the product-level first-difference

121

In the classic DID model, the first difference refers to the difference between before and after the experiment start time. The second difference refers to the difference between the treatment and control groups' first differences. Given the rich data structure at Amazon, such as the time series data of the business metric at product level, we can filter out some of the confounders that add noise to our estimates. We estimate the first-differences at the product level to help identify the heterogeneous effects of our treatment. The first-differences are estimated by taking the difference between the

126

127 product-level experimental period mean minus the product-level expectation based on the time trends
128 from the pre-experimental period which we refer to as β_i .

129 **3.2.1.2 Step 2: Filter out the outlier products**

130 During the experiment, there are always unobserved noises that could lead to extreme changes to the
131 business metric of a product. This leads to fat-tailed ATE distribution, under which basic averages or
132 regression without considering outliers is no longer the most efficient estimator (see [Athey et al. 2021](#)).
133 These extreme products add more noise than signal for our estimates. We define a rule to filter out those
134 extreme products outlined below.

135 First, we obtain a threshold percentage from the following equation:

$$141 P_{Th} = 1/(2\sqrt{N}) * 100$$

136 where N is the total number of products that are in the data. This cutoff is inspired work by [Vehtari et al.](#)
137 [\(2015\)](#). This threshold was validated through simulations and follows the intuition that as N increases,
138 the proportion of products in tails goes to zero while the number of tail products increases. We then drop
139 any products whose β_i falls outside the P_{Th} and $1 - P_{Th}$ quantiles of the β_i distribution of their treatment
140 group.

142 **3.2.1.3 Step 3: Product matching and second-difference**

143 Controlling for differences between treatment groups can improve the accuracy of estimates in
144 randomized controlled trials (RCTs) ([Deng et al 2013](#)). While imbalance can be mitigated with proper
145 randomization on the aggregate, differences between some metrics will naturally occur. We use Causal
146 Forests ([Wager and Athey 2018](#)) to control for differences between treatment groups. This allows us to
147 nonparametrically group products in the treatment group to similar products in the control group based
148 on specific product characteristics. Using causal forests, we can calculate estimated heterogeneous
149 treatment effects (HTE) for each product by comparing these similar products. However, we only report
150 the ATE to most lab owners to keep our results straightforward and easy to understand. We use each
151 products' average daily value of various financial metrics from the pre-experiment period as well as
152 other product characteristic information to group similar products in our Causal Forest.

153 **3.2.1.4 Step 4: Standard Error**

154 To estimate the standard error, we randomly sample products from our experimental population
155 (including outliers) with replacement. From this bootstrapped sample, we repeat our procedure, drop
156 outliers and estimate ATE using causal forests. We iterate K bootstraps to get the distribution of ATE
157 and then calculate the confidence bounds for the ATE of our important business metrics. This is called
158 randomization inference.

159 **3.2.2 HPTE Simulation**

160 To compare the effectiveness of our HPTE method compared to standard DID estimation, we used a
161 past experiment to simulate 200 random assignments of products to treatment and control groups. This is
162 called an A/A test. For each assignment, we estimate the ATE and standard error of the ATE. We
163 compare the average standard error and fraction of the time that we have a p-value less than 0.05
164 (indicating statistical significance). Because we are randomly assigning treatment, we expect the ATE to
165 be zero and the p-value to be less than 0.05 about 5% of the time. We report the average standard error,
166 the percentage of the time we have a P-value less than 0.05, and the standard deviation of our ATE

167 estimates in our simulation. We observe that HPTE estimates shrink the standard errors by about 30%
168 compared to DID.

169 **Table 1: HPTE Simulation Results**

| | Average SE | Pr(P-value<0.05) | SD of Sample ATE Estimates |
|-------------|------------|------------------|----------------------------|
| DID | 0.142 | 0.04 | 0.141 |
| HPTE | 0.104 | 0.03 | 0.087 |

170 **4 Spillover Effect**

171 Any A/B experiment consists of treatment and control groups. The treatment group is exposed to a new
172 policy while the control group is expected to be unaffected by the treatment. In the presence of
173 substitutable or complementary products in a pricing setting, the treatment can affect (spill to) the
174 controlled observations and bias the estimated treatment effect. This issue is known in the literature and
175 practice as spillover or interference. Such bias can result in significant deviations of the estimates from
176 true values and compromise the customer trust in pricing experiments. In this section, we aim to
177 characterize such bias using an exposure mapping technique (this method estimates the direct treatment
178 effect and indirect treatment effect due to spillovers), and reduce the bias using an effective cluster
179 randomization technique. In our numerical study, we observe a 30% reduction in bias on average when
180 using cluster randomization compared to the traditional DID approach with no spillover consideration
181 (referred to as Naïve approach henceforth).

182 In online pricing experiments, our main goal is to estimate the global treatment effect (i.e., the difference
183 in average outcomes when all units are exposed to treatment versus when all units are exposed to
184 control) and not the spillover effect. Yet, to motivate why and when the spillover bias problem should be
185 addressed, we study the measurement of the spillover effect. First, we identify the network of related
186 (i.e., substitutable or complementary) products. Next, we measure the spillover effect using an exposure
187 mapping technique. Finally, we address the spillover concerns using a balanced cluster randomization
188 and assess the performance of this approach in relation to a Naïve approach with no consideration for
189 spillovers. Throughout the paper to perform necessary numerical studies, we use past experiments run
190 on Amazon.com.






191 **4.1 Methodology**

192 **4.1.1 Building Network of Related Products**

193 In order to build the network of related products, we should start from a consideration set that identifies
194 which products can *potentially* be significant substitutes or complements of each other. We chose a
195 substitutable product service (SPS) at Amazon as a consideration set for this study which aggregates
196 substitute list from a variety of different models across Amazon and is available for more products
197 compared to the other sets available at Amazon. This substitutable product service identifies the
198 substitutes without considering price changes. In most pricing experiments, however, we are interested
199 in the substitution effect due to pricing changes for which an elasticity model that considers cross-price
200 elasticities is needed. We can find subsets of products within the consideration set that are significant
201 substitutes or complements of each other using cross-price elasticities. To build these cross-price
202 elasticity models, we use about a year of historical data for the products in the experiment. We use a
203 Poisson cross-price elasticity model that can be run for each experiment to identify relevant substitutes
204 using cross price elasticities from the set of possible substitutes identified by the substitute identification

205 service. The Poisson model seems to perform well in a few anecdotes that we checked. We present an
 206 example in Figure 5. Here, we find the related products to a stool. We observe that the Poisson model
 207 picks the stool with the same color and style from the consideration set as a substitute. The remainder of
 208 the consideration set, however, are items that are significantly different in terms of quantity, style, and
 209 price per unit. Hereafter, we use this model for identifying the network of related products.

210 **Figure 5: Identifying Substitutes Using a Poisson Cross-Price Elasticity Model**

| Item | Price | Color | Style | Cluster Model |
|---|---------|-----------------|-----------|---------------|
|  | \$65.46 | White & Natural | | |
|  | \$74 | White & Natural | Same | Poisson, SPS |
|  | \$112.8 | Brown | different | SPS |
|  | \$48.24 | Black | different | SPS |
|  | \$46.11 | Natural | different | SPS |

211

212 **4.1.2 Measuring the Spillover Effect**

213 One common approach to estimate the spillover effect is using the exposure models in combination with
 214 an inverse probability weighting (IPW) scheme like the Horowitz-Thompson (HT) estimator (Aronow et
 215 al. 2021). Here, we assume only direct peer spillover, in which case there are four possible exposures for
 216 a product:

- 217 1) Receiving the direct treatment only also called the isolated treatment effect (d_{10}),
- 218 2) Receiving direct treatment and indirect treatments through substitutes (d_{11}),
- 219 3) Receiving only indirect treatment through a substitute also called the spillover effect (d_{01}),
- 220 4) Receiving no treatment (d_{00}).

221 We next build a large enough sample using random draws of possible assignments to calculate the
 222 probabilities of exposures. Finally, we use Horvitz-Thompson (HT) estimate to calculate the treatment
 223 effects. For this analysis we focus on HT estimator that is known to have a lower bias compared to other
 224 IPW methods. We implemented this idea on the experiment using two months of pre-experiment and 4
 225 weeks of experimental data. We summarize the daily average aggregated treatment effects on QTY and
 226 the corresponding estimated standard deviations (SD) in Table 1.

227 **Table 2: Daily Average Aggregated Treatment Effects for the Experiment**

| Estimand† | Aggregated effect | SD |
|-----------------|-------------------|--------|
| Spillover | -404.43 | 123.72 |
| Isolated Direct | 936.32 | 389.15 |
| Naïve Treatment | 1420 | 70.84 |

228

229 Below are a few highlights:

- 230 • The experiment cannibalizes quantity sold in the control group.

- Ignoring spillover effect is inflating the treatment effect estimates under the Naïve DID approach with no spillover consideration.
- The Naïve approach aims at estimating the global effect which fails if SUTVA does not hold. This estimate in definition is different from the isolated direct effect. Estimating the global effect is not feasible when using the exposure mapping techniques. However, given the negative spillover in this study, one expects the global effect to be even less than the direct effect. Hence, we expect at least a 30% bias over-estimating the treatment effect when using the Naïve approach.

Given our interest in estimating the global treatment effects in pricing experiments, in the next section we focus on reducing spillover bias in global effect estimation using cluster randomization.

4.1.3 Addressing Spillover Concerns

One primary tool to address the spillover concern is cluster randomization, where clusters of substitute or compliment products are designed and same treatment is assigned to an entire cluster. This prevents spillover of treatment effect to the control through the related products if the related products are identified correctly. The downside of this approach is the larger variance and lower power as a result of smaller effective sample size since the number of clusters is less than the number of products.

4.1.3.1 Balancing treatment assignment and Power Analysis

Cluster randomization can lead to imbalance across the treatment assignments and hence introduce selection bias to our results. Thus, to improve the cluster randomization, we should factor in some cluster-level characteristics and cluster size. There are several techniques to achieve balance across the treatment and control groups upon cluster randomization including: 1) stratified block randomization, 2) clustered matched-pair randomization, and 3) constrained randomization. We selected constrained randomization after performing comparison studies across these methods. Under constrained randomization the following steps are taken to achieve balance:

- Specify important cluster-level covariates
- Simulate a large number of unique potential randomizations
- Choose a subset of randomizations where sufficient balance across covariates is achieved
- Randomly sample one randomization from this constrained space

We next performed power analysis. We observed that cluster randomization significantly reduces power compared to the product-level matched-pair randomization (status quo) as expected. This indicates that clustering is not suitable for low-powered experiments. We also observe that the constrained and matched-pair randomizations have comparable power results. Finally, the simulation-based power analysis results in 18-35% higher power while being more computationally intensive. More details are provided in Appendix 1.

4.1.3.2 Results

We next assess the performance of cluster randomization. We used a constrained randomization achieved in Section 4.1.3.1. We simulate the treatment and spillover effects and compare the Naïve approach to the Poisson cluster randomization. The bias and standard error (SE) trade-off highlights when one model is preferred to the other.

We simulated potential outcomes using the pre-experiment average for a financial performance metric at the product-level as the potential outcome for the control group. We next generate potential outcome for

272 products under direct treatment, indirect treatment, and both direct and indirect treatment using constant
 273 multipliers. We simulate a negative spillover effect by adjusting these multipliers. We created over 400
 274 of such multiplier vectors. Using the exposure map and the potential outcomes, we next calculated the
 275 observed outcomes and estimated the global treatment effect and standard error (SE) under the Naïve
 276 approach (with no account for the spillovers) and Poisson network cluster randomization. Additionally,
 277 we estimated the global treatment effect by generating one vector of potential outcomes for treatment
 278 (which was informed by whether the products were exposed to the direct/indirect treatment) and compare
 279 it to the potential outcomes for control. Table 3 illustrates a comparison of these estimates for a multiplier
 280 vector $(\alpha_{11}, \alpha_{10}, \alpha_{01}) = (1.2, 1.4, 0.85)$ when moderately significant treatment effect is detected under the
 281 Poisson model. Here we assume direct effect of 40% lift, spillover effect of 15% loss, and direct-indirect
 282 effect of 20% lift. Table 3 also includes the global treatment effect as the ground truth and bias
 283 improvement. Bias improvement measures the deviation of each estimate from the global treatment effect
 284 and illustrates the improvements in bias when using cluster randomization as opposed to the Naïve
 285 approach.

Table 3: Comparison of Naïve vs. cluster randomization

| α_{11} | α_{10} | α_{01} | Model | ATE (SE) |
|---------------|---------------|---------------|--------------------|---------------|
| 1.2 | 1.4 | 0.85 | Naïve | 51.23 (7.76) |
| | | | Poisson clustering | 31.15 (15.44) |
| | | | Global effect | 38.9 |
| | | | Bias improvement | 37% |

286
 287 Table 3 shows that this multiplier has a large direct treatment effect and the cluster randomization
 288 approach performs well in this case (recommending a launch based on the estimated standard errors). The
 289 Naïve approach in this case is highly inflated. In our numerical example of simulating 700+ treatment
 290 effects varying in a wide range, we observe:

- Using the Naïve approach can result in inflated estimates.
- The bias is highly sensitive to the simulated treatments.
- We observed an average of 30% reduction in bias by using cluster randomization compared to the Naïve approach.
- Standard deviation on average doubles when using the clustered randomization.
- Cluster randomization performs best when the treatment and spillover effect sizes are large and does not perform well when effect sizes are close to zero.

298 4.2 Spillover Summary

299 In this study, we performed an exposure mapping methodology to detect spillover and used a clustered
 300 constrained randomization to reduce spillover bias in our treatment effect estimates. We observed an
 301 average of 30% reduction in bias when using cluster randomization. However, cluster randomization does
 302 not perform well for low-powered experiments and can lead to noisy and unclear results. Thus, despite
 303 the presence of spillovers in all labs, certain categories of products may benefit more from the current
 304 proposed solution to address spillover. In particular, labs where high cannibalization is feared and larger
 305 sample sizes are available are most appropriate.

306
307
308
309
310
311
312
313
314

5 Conclusion

In pricing experiments, we want to make the lab owners learn as much as possible from the experiments they run. This involves making the results easy to understand and interpret for their use cases. This also involves making our estimates as accurate as possible. To help with this we improve experimental design by using things like Crossovers and improve our estimates by using HPTE. We also want to minimize bias which we do by studying and controlling for spillovers. Other avenues we continue to research to improve our experiments are improved randomization to ensure balanced treatment groups, cluster randomization to prevent spillovers, and HPTE enhancements to get more precise estimates of ATE.

315 **References**

316 Aronow, P., Eckles, D., Samii, C., & Zonszein, S. (2021). Spillover Effects in Experimental Data. In J.
 317 Druckman & D. Green (Eds.), *Advances in Experimental Political Science* (pp. 289-319).
 318 Cambridge: Cambridge University Press. doi:10.1017/9781108777919.021
 319 Athey, S., Bickel, P. J., Chen, A., Imbens, G., & Pollmann, M. (2021). *Semiparametric Estimation of*
 320 *Treatment Effects in Randomized Experiments* (No. w29242). National Bureau of Economic
 321 Research.
 322 Deng, Alex, et al. "Improving the sensitivity of online controlled experiments by utilizing pre
 323 experiment data." *Proceedings of the sixth ACM international conference on Web search and data*
 324 *mining*. 2013.
 325 Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2015). Pareto smoothed importance
 326 sampling. *arXiv preprint arXiv:1507.02646*.
 327 Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using
 328 random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.
 329

330 **Appendix 1 Power analysis with cluster randomization**

331 There are two different approaches in estimating the power that we study here: 1) an analytical closed-
 332 form expression considering a normality assumption, or 2) simulation-based power analysis. The closed-
 333 form expression is more computationally efficient while its assumptions are more difficult to justify. In
 334 particular, the closed-form expression is not valid for our constrained randomization and causal forest
 335 treatment effect estimation.

336 We next perform power calculations mentioned above using the experiment data. In randomizing the
 337 clusters, we balanced across the treatment and control groups on different financial metrics and the
 338 cluster size following the process described in Section 4.1.3. We consider a 6%, 8%, and 10% effect
 339 sizes (since this experiment was not well-powered, we selected larger effect sizes). Table 4 summarizes
 340 the power results.

341 **Table 4: Closed-Form and Simulation-Based Power Calculations**

| Effect size | CPPCP Power | | | | |
|-------------|--------------------------|-------------------------|-------------------------------|--------------|-------------------|
| | ASIN-level randomization | Clustered randomization | | | |
| | Matched-pair | Constrained | | Matched-pair | |
| | Closed-form | Closed-form | Simulation | Closed-form | Simulation |
| 6% | 0.3 | 0.17 | 0.2 [0.16, 0.23] [†] | 0.16 | 0.19 [0.15, 0.23] |
| 8% | 0.47 | 0.26 | 0.31 [0.27, 0.35] | 0.24 | 0.31 [0.27, 0.35] |
| 10% | 0.66 | 0.38 | 0.45 [0.4, 0.5] | 0.37 | 0.5 [0.45, 0.54] |

Notes: [†]The 95% confidence intervals are included for simulation-based power analysis.