Business Economics
https://doi.org/10.1057/s11369-023-00303-9

NABE
National Association for Business Economics

FOCUS ON INDUSTRIES AND MARKETS

Science of price experimentation at Amazon

Joe Cooprider[1] · Shima Nassiri[1]

Abstract
In order to improve prices at Amazon, we created Pricing Labs, a price experimentation platform. Since we do not price discriminate, we must run product-randomized experiments. We discuss how we randomize to prevent spillovers, run different experimental designs (i.e., crossovers) to improve precision, and control for demand trends and differences in treatment groups to get more precise treatment effect estimates.

Keywords Experimentation · Pricing · Multiple Randomization · Spillover Effect

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 website.

In this paper, we describe (1) what we do as scientists 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.

✉ Joe Cooprider
jccoopri@amazon.com

Shima Nassiri
shmnas@amazon.com

[1] Amazon Retail Pricing Science and Research, Seattle, USA

Published online: 21 January 2023

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 and is generally not used 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 Fig. 1, where the red cells are when a product is being treated.

Time-bound experiments are not always the best design for pricing experiments at Amazon. Our prices fluctuate based on a variety of different factors that can change over time (e.g., costs, promotions, prices at other stores, etc.). Some of these can change during our experiment regardless

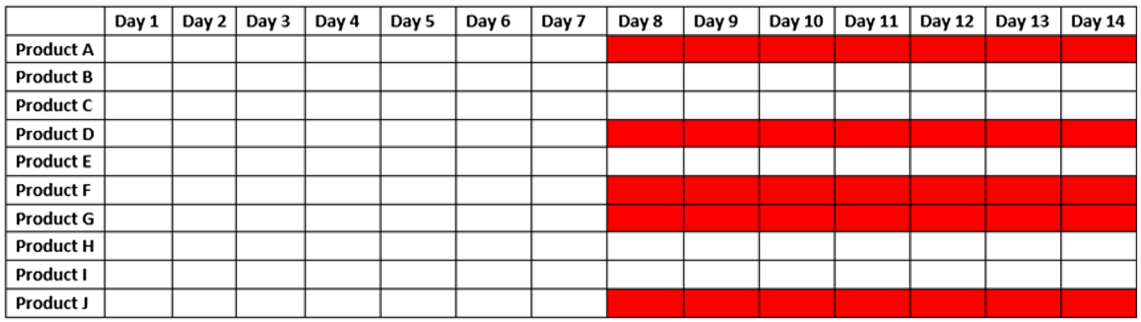


Fig. 1 Time-bound experimental design

of what policy we are testing. Changes in factors that determine our prices during the experiment can change the experiment population. We use trigger-based experiments for these cases. A trigger-based experiment is when a product is only in the experimental analysis after a "trigger" is met. This means that only a subset of the original experiment population is analyzed. Once a product is triggered, it enters the experiment regardless of the treatment group it belongs to. If the product is in the treated group, the new policy will be applied to it after being triggered, while products in the control group continue with the existing policy. When products get triggered, we consider them as triggered until the end of the experiment.

Below is an example of a trigger-based experiment. Red cells are treated experimental periods and green cells are control experimental periods. Suppose the trigger is a product being put on promotion at another store. Once a product is triggered (another store puts it on promotion), it enters the experiment and is assigned to either treatment or control. In this example, products A and E are on promotion at another store on day 8 (the first day of the experiment). Products B, F, G, I and J are put on promotion at another store during the experiment and are added to the analysis the day that they are put on promotion. Note that in this example, three products (C, D, and H) are never triggered and are left out of the analysis (Fig. 2).

3 Improving precision

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

3.1 Experimental design

Switchbacks are a common tool to improve precision and power in experiments. For this paper, we define switchbacks as when treatment varies across products and time during our experiment. Switchbacks generally occur when the treatment turns on or off multiple times for each product in

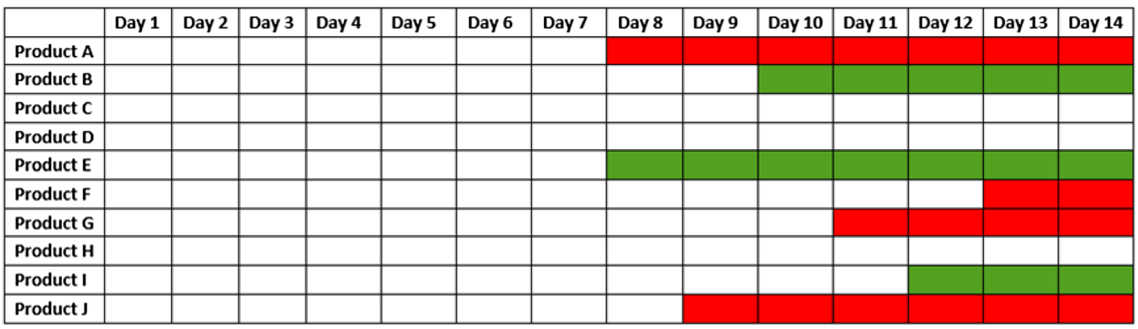


Fig. 2 Trigger-based experimental design

the experiment. For the time-bound example above, products would switch between the new machine learning (ML) algorithm price and the old ML algorithm price multiple times throughout the experiment. This is beneficial for many reasons: it exposes more products to treatment since each product can be treated during the experiment, it increases the variation of when the treatment is applied to each product since the start of the treatment is different for different products, and it provides a more effective counterfactual since each product has both treatment and control periods during the experiment.

In this setting, since the treatments start on different days for different products, it allows us to separate demand shifts across Amazon or among groups of products from the treatment more effectively. Further, we will have days within the experiment where each product is not treated, which will provide a more effective counterfactual than the control group or the treated group before the experiment began, the standard counterfactual periods in normal *A*/*B* tests. Below is an example of a switchback design called random days which randomly assigns each product-day to either treatment or control. Random days experiments can shrink standard errors by about 60% (Fig. 3).

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

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

Below is an example of crossover experimental design where week 7 is the start of the experiment and week 10 is the start of the second half of the experiment. Weeks 7 and 10 are blacked out because they are dropped from our analysis, but have the same treatment status as weeks 8 and 11, respectively. Consider we are comparing two ML algorithms. Group A would be priced with the new ML algorithm from week 7 to week and the old algorithm from week 10 to week 12. Group B would be priced with the old algorithm for weeks 7–9 and the new algorithm for weeks 10–12. Our analysis would not include weeks 7 and

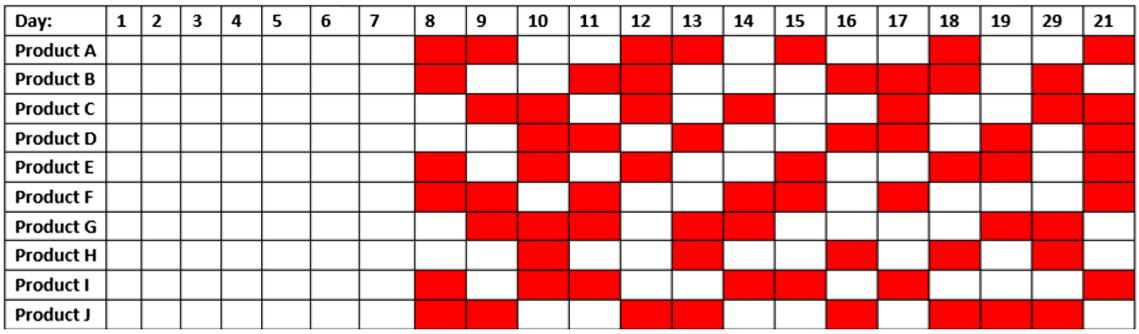


Fig. 3 Random days experimental design

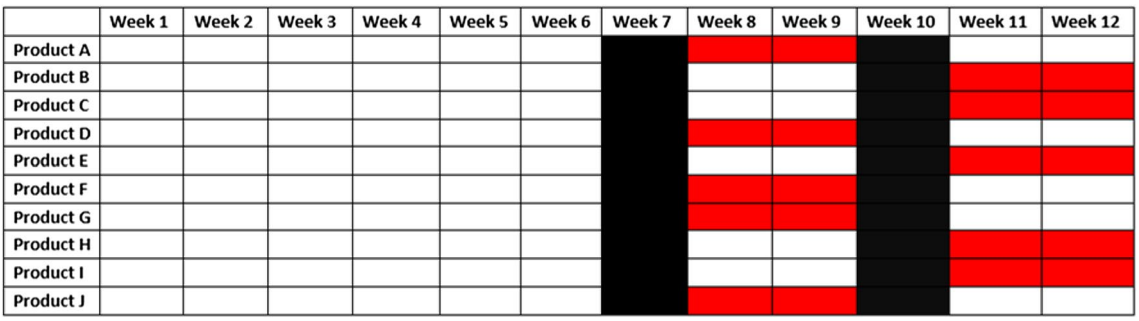


Fig. 4 Crossover experimental design

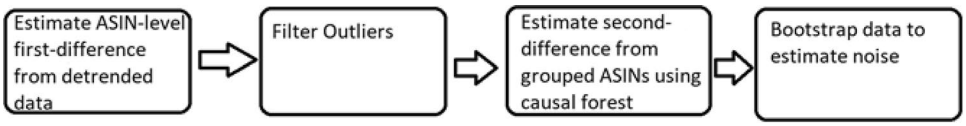
10 because those effects could be biased by carryover effect from the prices in the previous week (Fig. 4).

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.

3.2 Heterogeneous panel treatment effect (HPTE)

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).

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:



3.2.1 Methodology

3.2.1.1 Step 1: Estimate the product-level first-difference 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 product-level experimental period mean minus the product-level expectation based on the time trends from the pre-experimental period which we refer to as β_i.

3.2.1.2 Step 2: Filter out the outlier products During the experiment, there are always unobserved noises that could lead to extreme changes to the business metric of a product. This leads to fat-tailed ATE distribution, under which basic averages or regression without considering outliers is no longer the most efficient estimator (Athey et al. 2021). These extreme products add more noise than signal for our estimates. We define a rule to filter out those extreme products outlined below.

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

$$P_{\text{Th}} = 1 \Big/ \left(2\sqrt{N}\right) * 100,$$

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

3.2.1.3 Step 3: Product matching and second difference Controlling for differences between treatment groups can improve the accuracy of estimates in randomized controlled trials (RCTs) (Deng et al. 2013). While imbalance can be mitigated with proper randomization on the aggregate, differences between some metrics will naturally occur. We use Causal Forests (Wager and Athey 2018) to control for differences between treatment groups. This allows us to nonparametrically group products in the treatment group to similar products in the control group based on specific product characteristics. Using causal forests, we can calculate estimated heterogeneous treatment effects (HTE) for each product by comparing these similar products. However, we only report the ATE to most lab owners to keep our results straightforward and easy to understand. We use each products' average daily value of various financial metrics from the pre-experiment period as well as other product characteristic information to group similar products in our Causal Forest.

3.2.1.4 Step 4: Standard error To estimate the standard error, we randomly sample products from our experimental

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

population (including outliers) with replacement. From this bootstrapped sample, we repeat our procedure, drop outliers and estimate ATE using causal forests. We iterate K bootstraps to get the distribution of ATE and then calculate the confidence bounds for the ATE of our important business metrics. This is called randomization inference.

3.2.2 HPTE simulation

To compare the effectiveness of our HPTE method compared to standard DID estimation, we used a past experiment to simulate 200 random assignments of products to treatment and control groups. This is called an *A/A* test. For each assignment, we estimate the ATE and standard error of the ATE. We compare the average standard error and fraction of the time that we have a p-value less than 0.05 (indicating statistical significance). Because we are randomly assigning treatment, we expect the ATE to be zero and the p-value to be less than 0.05 about 5% of the time. We report the average standard error, the percentage of the time we have a p-value less than 0.05, and the standard deviation of our ATE estimates in our simulation. We observe that HPTE estimates shrink the standard errors by about 30% compared to DID (Table 1).

4 Spillover effect

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

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

4.1 Methodology

4.1.1 Building the network of related products

In order to build the network of related products, we should start from a consideration set that identifies which products can *potentially* be significant substitutes or complements of each other. We chose a substitutable product service (SPS) at Amazon as a consideration set for this study, which aggregates a substitute list from a variety of different models across Amazon and is available for more products compared to the other sets available at Amazon. This substitutable product service identifies the substitutes without considering price changes. In most pricing experiments, however, we are interested in the substitution effect due to

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

Fig. 5 Identifying substitutes using a Poisson cross-price elasticity model

pricing changes for which an elasticity model that considers cross-price elasticities is needed. We can find subsets of products within the consideration set that are significant substitutes or complements of each other using cross-price elasticities. To build these cross-price elasticity models, we use about a year of historical data for the products in the experiment. We use a Poisson cross-price elasticity model that can be run for each experiment to identify relevant substitutes using cross-price elasticities from the set of possible substitutes identified by the substitute identification service. The Poisson model seems to perform well in a few anecdotes that we checked. We present an example in Fig. 5. Here, we find the related products to a stool. We observe that the Poisson model picks the stool with the same color and style from the consideration set as a substitute. The remainder of the consideration set, however, are items that are significantly different in terms of quantity, style, and price per unit. Hereafter, we use this model for identifying the network of related products.

4.1.2 Measuring the spillover effect

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

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

We next build a large enough sample using random draws of possible assignments to calculate the probabilities of exposures. Finally, we use Horviz–Thompson (HT) estimates to calculate the treatment effects. For this analysis we focus on HT estimator that is known to have a lower bias compared to other IPW methods. We implemented this idea on the experiment using two months of pre-experiment and 4 weeks of experimental data. We summarize the daily average aggregated treatment effects on QTY and the corresponding estimated standard deviations (SD) in Table 2.

Below are a few highlights:

- 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 the stable unit treatment value assumption (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.

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

4.1.3 Addressing spillover concerns

One primary tool to address the spillover concern is cluster randomization, where clusters of substitute or complement 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:

i. Specify important cluster-level covariates

ii. Simulate a large number of unique potential randomizations
iii. Choose a subset of randomizations where sufficient balance across covariates is achieved
iv. 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 the Appendix.

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

Table 3 shows that this multiplier has a large direct treatment effect and the cluster randomization approach performs well in this case (recommending a launch based on the estimated standard errors). The Naïve approach in this case is highly inflated. In our numerical example of simulating 700 + treatment 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 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.

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%

4.2 Spillover summary

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

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 using things like Crossovers

Table 4 Closed-form and simulation-based power calculations

Effect size (%)	ASIN-level randomization matched-pair closed-form	CPPCP power			
		Clustered randomization			
		Constrained		Matched-pair	
		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]

and improve our estimates 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.

Appendix: Power analysis with cluster randomization

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

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

Funding Funding was provided by Amazon.

References

Aronow, Peter M., Dean Eckles, Cyrus Samii, and Stephanie Zonszein. 2021. Spillover Effects in Experimental Data. In *Advances in Experimental Political Science*)J. Druckman and D. Green, editors). Cambridge: Cambridge University Press.

Athey, Susan., Peter J. Bickel, Alloy Chen, Guido W. Imbens, and Michael Pollmann. 2021. *Semiparametric Estimation of Treatment Effects in Randomized Experiments*. National Bureau of Economic Research Working Paper no. 29242.

Deng, Alex, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre experiment data. *Proceedings of the sixth ACM international conference on Web search and data mining*.

Vehtari, Aki., Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. 2015. *Pareto smoothed importance sampling*. http://arxiv.org/abs/1507.02646.

Wager, Stefan, and Susan Athey. 2018. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association* 113 (523): 1228–1242.

Joe Cooprider graduated with a Ph.D. in Economics from Boston College in 2020. His research focused on micro-econometric methods including estimating heterogeneous demand for various products (i.e., junk food and soda). Joe started work for the Consumer Pricing Research team at Amazon in 2020 where he started working on Pricing Labs and other econometric challenges of pricing.

Shima Nassiri has been a research scientist on the Amazon Retail Science and Research team since September 2021. She works on projects related to pricing experimentation. Her latest projects are on treatment heterogeneity detection, reduction of bias through cluster randomization, and subgroup analysis. She received her PhD in Operations Management from the University of Washington. Prior to joining Amazon, she was as an assistant professor at the University of Michigan Ross School of Business where her research was mainly related to healthcare economics. She has publications in top operations management journals like *Management Science* and *Manufacturing & Service Operations Management*.