The Price is Right: Removing A/B Test Bias in a Marketplace of Expirable Goods

Thu Le
Airbnb
San Francisco, CA, USA
thu.le@airbnb.com

Alex Deng
Airbnb
Seattle, Washington, USA
alex.deng@airbnb.com

ABSTRACT

Pricing Guidance tools at Airbnb aim to help hosts maximize the earning for each night of stay. For a given listing, the earning-maximization price point of a night can vary greatly with lead-day - the number of days from now until the night of stay. This introduces systematic bias in running marketplace A/B tests to compare the performances of two pricing strategies. Lead-day bias can cause the short-term experiment result to move in the *opposite* direction to the long-term impact, possibly leading to the suboptimal business decision and customer dissatisfaction. We propose an efficient experimentation approach that corrects for the bias, minimizes the possible negative impact of experimenting, and greatly accelerates the R&D cycle. This paper is the first of its kind to lays out the theoretical framework along with the real-world example that demonstrates the magnitude of the bias. It serves as a conversation starter for such insidious type of experimentation bias that is likely present in other marketplaces of expirable goods such as vacation nights, car rentals, and airline tickets, concert passes, or ride-hailings.

CCS CONCEPTS

• **Mathematics of computing** → **Probabilistic inference problems**; • **Applied computing** → **E-commerce infrastructure**.

KEYWORDS

A/B Testing, Marketplace Experimentation, Online Controlled Experiment, Dynamic Pricing, Heterogeneous Treatment Effect

ACM Reference Format:

Thu Le and Alex Deng. 2023. The Price is Right: Removing A/B Test Bias in a Marketplace of Expirable Goods. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3583780.3615502

1 INTRODUCTION

Online controlled experiments (e.g. A/B tests) have been the gold standard in understanding the impact of a potential feature change[23, 26, 27]. Pricing Guidance at Airbnb is no exception. Airbnb does not control the nightly prices set by hosts. To help them respond to the travel trends and maximize earnings, we offer tools that provide hosts with tips or suggestions for pricing their listings. We constantly improve our recommendation engine, leveraging A/B testing as guidance throughout the process. In many of the experiments, we often observe a strong "novelty" effect [24, 27, 42], in which, the revenue impact measured in the experiment often trends strongly either in the positive and negative direction before changing its course and stabilize at a certain level, but only after months of running. Fig. 1 demonstrates an example of such novelty effect.

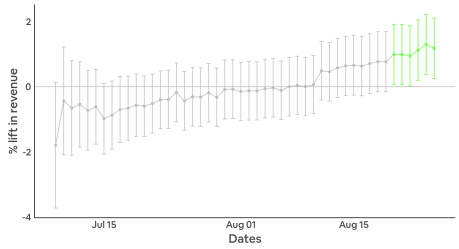

Figure 1: The puzzling outcome of a pricing experiment. The revenue metric shown here is aggregated over all the bookings since the start of the experiment to the current date. Very often, this metric trends strongly in once direction, negative in this case, before changing course, and it could take months before metric converges.

Most of the feature changes in these experiments happen in the backend, primarily in recommended prices to hosts, with virtually no interruption in the user-facing components. Thus, the classic novelty effect [17, 18, 33], in which, users tend to find the new interface experience either more fascinating or more disorienting for a short period of time, does not apply. Instead, this is the systematic result of the interaction between the pricing strategy and lead-day, i.e. the number of days between now and the night of stay that we provide the recommendation for. This lead-day bias decouples the short-term observation of the treatment effect from its true long-term impact, and the implied revenue gain or loss at stake could be significant for a platform with millions of active listings.

Aside from seasonality, lead-day is a major component in pricing for perishable goods such as vacation nights, concert tickets, airplane tickets etc. Sellers tend to update the prices accordingly: price high at the beginning and lower it later if the good is not sold, or give out a discount to the early birds to ensure the sale. To help hosts maximize revenues, our pricing recommendation models are heavily lead-day dependent: the tools provide updated suggestions for nightly prices daily to account for how much relevant inventory is left on the platform and how close the booking dates are. Thus, the revenue-maximization price point for a summer trip can vary significantly depending on the time of booking, taking the full advantage of the lead-day dynamic. Fig. 2 demonstrates the lead-day dynamic of a sample price recommendation. More details on our price recommendation system architect can be found in [43].

The standard way of conducting A/B tests for pricing recommendations is to randomize the pool of listings into the treatment and control groups. For the control group, all the pricing recommendations are generated by the model in production as usual while for the treatment group, we immediately switch the recommendations

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615502

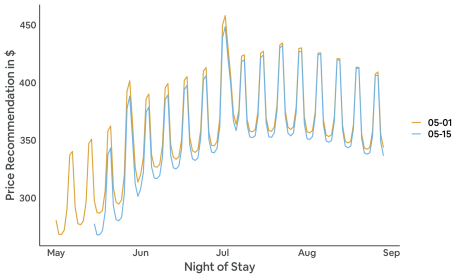

Figure 2: The orange line is the price recommendations for summer nights at the beginning of May, and the blue line price recommendations two weeks later. The blue line is significantly below orange line for upcoming nights.

to the new test model post randomization. As a result, the upcoming nights in the treatment group would be priced under two separate and uncoordinated models, production one earlier and development one later. We call this Mixed Treatment Effect and it will mainly bias against the performance of the nights that are close to the experiment start date in the treatment group. Another bias is introduced through accounting: in A/B test, our revenue metric is recognized on the day of the booking regardless of the check-in date, and it will favor the pricing strategy that sells the night at a faster rate, but not necessarily the one that fetches the highest final price. We call this Fill Rate Bias.

This paper is the first of its kind to discuss the phenomenon of how a pricing algorithm for perishable items, vacation nights in this case, can interact with lead-day, making the short-term experiment point to the *opposite* direction of the long-term impact in many cases. The major contributions of this paper include:

- Present and bring awareness of lead-day bias that can cause the short-term effect to move in the opposite direction of the long-term effect.
- Identify two sources of bias in the marketplace experiments related to lead-day.
- With the understanding of two biases, propose a three-prong approach to correct for bias and accelerate the experimentation process by:
  (1) Limiting treatment roll-out to nights that are far enough lead-day,
  (2) Smart-overlapping of the experiments through stack-up scheduling,
  (3) Applying a Heterogeneous Treatment Effect (HTE)-remixed estimator to predict the long-term effect.
- Verify with real-world experiments. Our approach is applicable to a broad set of pricing experiments with expirable/perishable goods.

Lead-day bias is not Airbnb-specific but it applies to all the marketplaces of expirable goods such as travel reservations, hail-ridings, airline tickets, or concert passes. We suspect that many other marketplace companies might have missed this interference in running A/B tests, which could bias the decision-making process and affect the consumers as a result. [22] details the design of the experimentation platform at a big travel platform without mentioning the issue with lead-day bias and how the platform can accommodate it. This paper serves as the conversation starter about such an important and insidious bias in A/B testing. The solution proposed by the paper, while working reasonably well in our case, can be iteratively improved through the contribution of the industry experts once it is presented.

The rest of the paper is organized as the following: Section 2 highlights past researches from the industry as well as academia on the topic of short vs. long-term treatment effect as well as other challenges of marketplace experimentation. We then dissect lead-day bias into two components in Section 3 and illustrate them by a real pricing experiment in Section 3.3. We propose the corrections for both types of lead-day bias in Section 4. Section 5 demonstrates our method using the same real pricing experiment.

For readers new to the subject, below is a few essential terminologies to understand the marketplace experiment setup at Airbnb:

- **Listing**: This is a unit of vacation rental at Airbnb. Once a listing is onboarded to the platform, hosts can start setting the availability and price for each future night. Hosts can either set a flat price for all the nights or a personalized the price for each night. Our Pricing Guidance tools can also provide hosts with the nightly price suggestions.
- **Night**: This is referred to the actual date of a stay, and it is forward-looking. Guests can book a night that is up to two years in advance. Most of the bookings, however, are for much closer check-in dates.
- **Lead-day**: The number of days from the current date to the night that we want to help hosts price to sell. A same-day booking will have a lead-day of 0, and the next day a lead-day of -1. Both are examples of short lead-day bookings.
- **Booking Cycle**: This is referred to the period, in which, a night could be booked. A night is typically booked from 120 days ahead until same-day booking, or from -120 to 0 lead-day.
- **Realized Night**: A night that has passed, and we no longer could sell it. We can also refer to the night as expired.
- **Realized Booking**: A booking with check-in date in the past.
- **Realized Revenue**: Revenue from realized bookings.

2 RELATED WORKS

Divergence between short vs. long-term treatment effect both experimentally and observationally is not new. Industry leaders in A/B tests have also studied extensively about the novelty effect [9, 13, 18, 24, 25, 33]. They, however, mostly deal with the user learning rate [37], in which, users' level of engagement changes over time depending on their familiarity with the treatment. This novelty vs. primacy effect issue is also well-known in psychology [1, 6, 21, 31, 32, 36], medicine [4], sport science [16, 20], sociology [38], and healthcare [30]. The technical solution is usually top-down by estimating the user learning rate by comparing the Average Treatment Effect (ATE) among different experiment periods. Lead-day bias is slightly different, in which, it is the result of the interaction between the treatment and lead-day, and a more targeted solution of using lead-day HTE-remixed estimator is leveraged in this paper.

In academia, [3] discusses in detail the use of short-term surrogate index to project the long-term impact. In order for a surrogate index to work, long-term outcome needs to be independent of the treatment, conditional on the full set of surrogates. This is not feasible given the complex interaction between pricing and market dynamics at Airbnb. It also uses a single point-in-time set of surrogate metrics to predict long-term, which does not capture the time dynamism. Another promising approach is to combine both short-term experiment data with long-term observational data suggested by [2] to estimate the final outcome. In Airbnb case, lead-day bias can make the short-term result move in the opposite direction of the long-term, and it is risky to launch the test variant based on a short-running experiment and observationally estimate the long-term impact. Other directions and improvements in developing surrogacy for A/B testing by

[5, 10, 39] are promising, but most assume that the treatment remains stable over time. In the case of dynamic pricing, it is not guaranteed that the treatment will remain stable as the pricing outputs can change dramatically over time depending on the market condition.

Aside from short vs. long-term treatment effect issue, marketplace experiments can be prone to several other types of bias. Cannibalization, for one, where the treatment applied on one experiment arm can affect the outcome in another arm, is one of the most outstanding and well-known issues. There have been several attempts to understand and correct this bias[7, 19, 28, 29]. Relatedly, network effect is another serious issue, and there exists a large body of works dedicated to understanding and addressing it in the context of running A/B tests[8, 14, 34, 42]. Lastly, companies and platforms have invested heavily into improving test sensitivity as we very often deal with long-tailed test populations[12, 41].

Despite all of the above efforts, lead-day bias remains an elusive topic, and as far as our understanding, no previous researches in either industry or academia have tangentially mentioned and attempted to address this kind of experimentation interference. With this paper, we hope to invite more discussions on how to solve such insidious and widespread inference bias.

3 DISSECTING LEAD-DAY BIAS

For simplicity, in this scenario, our goal is to find the pricing strategy that maximizes earning for May 1st night and share that information with hosts. This scenario also assumes that the listings are homogeneous and should be priced with the same pricing strategy. For illustration, say we have a flat pricing policy in production: it prices the night is the same regardless of the lead time. The competing algorithm prices the night more dynamically: it prices the night relatively more expensive at the early book date and starts dropping the price at a later date if the night remains unsold. Fig. 3 demonstrates the two pricing strategies that we want to compare against each other.

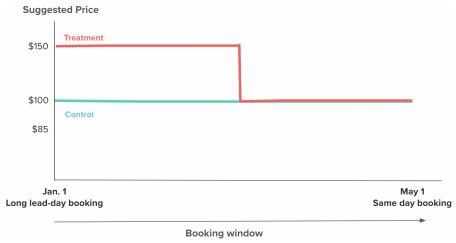

Figure 3: Two different pricing strategies that vary price by lead-day for May 1st night.

We conduct the A/B test and randomize our pool of listings into the control group which will stay under the production model and the treatment group which will be switched to the test model immediately post-randomization. Assume we launch the experiment on February 1st and begin observing the outcomes after 3 weeks to make the launch decision. All the bookings made during this period for May 1st check-in date will be considered in the analysis. Under this setup, the final experiment result will only capture a portion of the booking windows as shown in Fig. 4, and it introduces biases.

3.1 Mixed Treatment Effect

The first type of bias is due to the listings in the treatment group being priced under two distinct pricing strategies. Prior to the start of the experiment, a sizable portion of the inventory had been sold under the production model. The remaining of the inventories was switched immediately to the test model after the experiment starts.

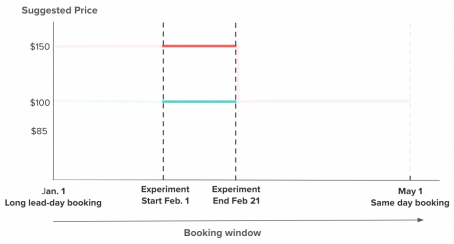

Figure 4: The A/B test only captures a portion of the booking window.

In this scenario, since the production model priced the night cheaper, many desirable listings might have already been sold before February 1st, leaving the test model with little room left to raise the price. This discontinuity in the pricing strategy works against the listings in the treatment group, and we call this Mixed Treatment Effect. See Fig. 5 for the demonstration of this bias.

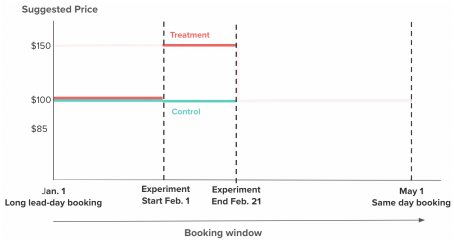

Figure 5: Mixed Treatment Effect causes bias against listings in the treatment group due to the discontinuity in the pricing strategy.

3.2 Fill Rate Bias

The second type of bias is due to how we account for the booking metrics. As mentioned above, we credit any booking made during the runtime of the experiment toward the final result. As in this case, our experiment ends before May 1st, the experiment result may not be reflective of how much revenue each algorithm could have generated by the night's expiration date. Fig. 6 demonstrates this accounting issue: During the experiment runtime, our production model prices May 1st night cheaper, thus, it could sell faster. Hypothetically, we could sell three nights in the control group for $100/night, and the final revenue is $300 by the end of the experiment. In the treatment group, we price the night at $150, and only manage to sell one night, and the total take home is $150. This leaves us with -50% drop in revenue for the treatment group, and the test model is a no-go. Had we let the experiment to run until May 1st, we could have seen the treatment group dropping the prices to $100/night. We could have managed to sell additional 3 nights in the treatment group at the lower price, and the realized revenue is $450 while the control group only sells one more night at the flat price of $100/night. Thus, the final outcome could have been $450 for the treatment group vs. $400 for the control group. This is +12.5% lift, a huge win for the test model which we could have missed out. We call this Fill Rate Bias due to it favors the pricing strategy that fills the room faster, not necessarily at higher revenue.

One salient point is that in practice, the final experiment outcome is aggregated over many nights together, and this makes the debiasing work even more challenging and complex. In the next section, we will go through the result of a real Smart Pricing experiment that we ran between 2021-2022 to demonstrate this point.

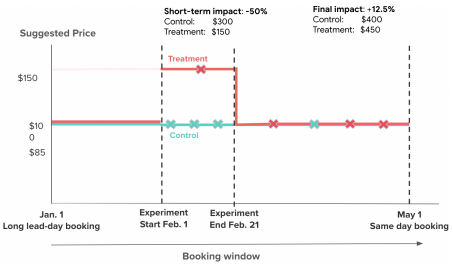

Figure 6: Fill Rate will favor the pricing strategy that fills the room at a faster rate and not necessarily leads to the highest revenue possible.

3.3 Empirical evidence from a real pricing experiment

We demonstrate both types of bias using data from a real experiment that ran between December 2021 and April 2022. For this experiment, we tested out a new revenue-maximization strategy layer for listings that were enrolled in Airbnb Smart Pricing. In treatment group, the average recommended price was about 2% higher than the control for longer lead-day bookings and dropped aggressively as the check-in dates draw closer.

Since the listings in the treatment group were listed at higher prices, for the early bookers, we will expect the nights to be sold at slower rates, and revenue might only catch up at the approaching of the check-in date. At the same time, due to Mixed Treatment Effect, the earlier nights at the start of the experiment might not perform as well as the later nights which enjoyed more continuous pricing strategies in the treatment group. Figure 7 provides the cohort view, in which, we line up the revenue lift by lead-day for a fixed group of nights. Horizontally, as the check-in date gets closer, the revenue lift gradually improves due to Fill Rate Bias winding down. Vertically, the further away the nights are to the experiment start date, the better the test model performs due to Mixed Treatment Effect early on.

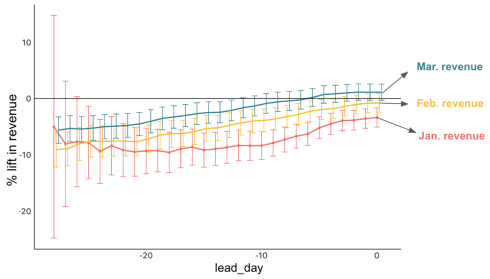

Figure 7: Revenue lift by lead-day for nights of the first week of January, February, and March nights. Due to Fill Rate Bias, revenue lift gets better as lead-day moves closer to zero for all the nights. Also, the further the nights are to the start of the experiment, the better it performs due to the Mixed Treatment Effect is stronger January nights than February nights than March nights.

In summary, in order to help hosts optimize for revenue, our pricing recommendation models are heavily lead-day dependent. This introduces lead-day bias when running an A/B test: Mixed Treatment Effect biases against listings in the treatment group in the earlier nights due to them being priced under two distinct models while Fill Rate will favor the pricing model that sells the nights at faster rates. The next section will discuss on a combined proposal to correct these biases by redesigning our experiment and post-analysis adjustment.

4 CORRECTING LEAD-DAY BIAS

4.1 Mitigate Mixed Treatment Effect through design

Under the default A/B test framework, we would roll out the test model on the treatment group immediately after randomization. In our case, it causes the discontinuity in pricing strategy for those nights that are close to the experiment start date in the treatment group. As shown in the Fig. 7 above, January and February nights are heavily impacted by Mixed Treatment Effect. Thus, instead of rolling out the test model to all the nights, it is better for us to limit the test model to nights that are sufficiently far away from the experiment start date and limit our analysis to the bookings on these nights to mitigate the Mixed Treatment Effect.

Airbnb's main pricing model usually prices nights 120 days ahead, and the majority of the bookings are captured within this maximum lead-day. The Mixed Treatment Effect would be completely removed if we limit rolling out the test model to nights that are 120 days ahead. In practice, the 120-day period could be significantly shortened as the bulk of the bookings happens within two months in advance. For each experiment, depending on the seasonality, we could decide ahead of time on how far ahead we need to start rolling out the model. For example, during the high-demand season of summer, nights tend to be booked far in advance. For experiments that start in May, it is better to bypass summer nights and only roll out the test model starting for September nights. In this experiment, we could have been better off rolling out the test model for March nights despite the experiment's start date in December.

4.2 Smart-overlapping of experiments

Pricing experiments tend to have low sensitivity due to the high degree of heterogeneity among Airnbb listings. We usually have to dedicate the entire experiment bandwidth to one experiment at a time, and two pricing experiments, despite being orthogonal, need to be run sequentially. We can only run a handful of experiments in a year, which critically slows down our pace of innovations. The limited roll-out approach proposed above ensures that the experiment captures the treatment effect for the entire booking cycle, and the analysis is restricted to one specific set of nights. This also frees up experiment bandwidth, allowing us to launch other experiments for non-overlapping sets of nights. For example, while running the current experiment that started in December with treatment and impact analysis limited to March nights, we could fit in another experiment that starts in January with treatment and impact analysis limited to April nights, and so on. Fig. 8 demonstrates this experiment stack-up strategy. Under this proposal, in theory, we could run up to nine experiments in a year instead of three, and significantly accelerates the R&D process. Admittedly, there will be some spillover where trips are booked across the months, and we would need some rest intervals among experiments. The pacing between two consecutive experiments will need to be adjusted depending on seasonality as mentioned previously. This is a small price to pay in exchange for the nimbleness in testing and iterating. In the next section, we will go into details of another technical workaround that can shorten the runtime of a pricing experiment and allow for even more experiments to fit into our testing schedule.

4.3 Predict long-term final effect from short-term experiment with HTE remixing

Due to Fill Rate Bias, evaluating a new pricing policy normally requires running an experiment long enough such that most nights in the experiment are realized. At the same time, to mitigate the Mix Treatment Effect, we will roll out and experiment on nights sufficiently far away from the experiment start. This means the experiment needs to run for at least four months for conclusion! In practice, this will significantly lengthen the R&D cycle. Instead of passively running a long-term experiment to avoid the Fill Rate Bias,

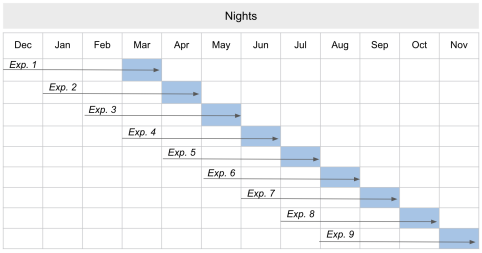

Figure 8: Ideal roll-out schedule of pricing experiments. Highlighted cells indicated nights that the treatment and analysis are limited to for each experiment.

we propose a method to adjust lead-day bias proactively so that we can predict the long-term effect from a shorter-term experiment. In particular, instead of running a four-month experiment, we can get an early read on the experiment as soon as two months post-launch. Aside from operational improvement, this is acutely important in preventing the negative impact of the test model on revenue.

Let i and j be the indices for nights and lead-days respectively and l for listings. Following the last section, we only consider lead-day within 120 days so $j=-120,\ldots,0$. Let r_{ij}^l be the total revenue for the night i of a listing l booked j days ahead. Averaging across all listings in a treatment group $g=t,c$ we get

$$m_{ij}^g = \frac{1}{n_g}\left[\sum_{l=1}^{n_g} r_{ij}^l\right].$$

For each i,j, we can compute the lift of revenue for night i booked j days ahead as

$$\text{LIFT}_{ij} = \frac{m_{ij}^t - m_{ij}^c}{m_{ij}^c},$$

and lift of revenue for nights i from all lead-day j as

$$\text{LIFT}_i = \frac{\sum_{j=-120}^0 m_{ij}^t - \sum_{j=-120}^0 m_{ij}^c}{\sum_{j=-120}^0 m_{ij}^c}.$$

Define $w_{ij} = \frac{m_{ij}^c}{\sum_{j=-120}^0 m_{ij}^c}$ be the proportion of revenue booked j days ahead in the *control* group. In other words, for each i, w_{ij} represents how fast listings were booked as lead-day j approaches 0 in the current control pricing strategy. With this notation,

$$\sum_{j=-120}^0 w_{ij}\times\text{LIFT}_{ij} = \sum_{j=-120}^0 \frac{m_{ij}^c}{\sum_{j=-120}^0 m_{ij}^c}\times\frac{m_{ij}^t - m_{ij}^c}{m_{ij}^c}$$
$$= \frac{\sum_{j=-120}^0 m_{ij}^t - \sum_{j=-120}^0 m_{ij}^c}{\sum_{j=-120}^0 m_{ij}^c}.$$

Therefore,

$$\text{LIFT}_i = \sum_{j=-120}^0 w_{ij}\times\text{LIFT}_{ij}. \quad (1)$$

Equation (1) says for any given night i the lift of revenue is a weighted average of observed lift of revenue spread across different lead-day j. This sounds obvious and we are just segmenting the revenue metric by lead-day. But there is a fundamental difference: the weights w_{ij} are defined using the control group only, and the revenue mix from treatment policy can be very different from the control!

Two critical insights from Equation (1) leading to our proposed correction method are:

(1) LIFT_{ij} is simply a function of lead-day j and not related to i if we assume the Fill Rate Bias manifest itself through lead-day. In a short-term experiment, we may not be able to observe LIFT_{ij} for j close to 0, but when the trend is strong, a regression model can be used to extrapolate the trend.

(2) w_{ij} may seem to also depend on whether we run the experiment long enough for all nights to be realized. However, because it is a property of the current control policy, we can estimate them using past observational data.

The derivation above entails the following result.

Theorem 1. *Assume* $\text{E}(LIFT_{ij}) = f(j)$ *for a function* f *of lead-day* j, *i.e.*

$$LIFT_{ij} = f(j) + \epsilon_{ij}, \text{E}(\epsilon_{ij}) = 0. \quad (2)$$

If estimates $\widehat{f}(j)$ *are asymptotically unbiased for* $f(j)$, *then*

$$\sum_{j=-120}^0 w_{ij}\times\widehat{f}(j) \quad (3)$$

is asymptotically unbiased for $\text{E}(LIFT_i)$.

Modeling $\text{E}(LIFT_{ij})$ as a function of lead-day i is a special type of heterogeneous treatment effect (HTE) [40]. In our case, we have point estimates of $LIFT_{ij}$ as well as their variances using the Delta method [11]

$$Var(\text{LIFT}_{ij}) = \frac{1}{(m_{ij}^c)^2}\left((\sigma_{ij}^t)^2 + \frac{(m_{ij}^t)^2}{(m_{ij}^c)^2}(\sigma_{ij}^c)^2\right) \quad (4)$$

where $(\sigma_{ij}^g)^2$ is the sample variance of revenues generated by listings in the group g for bookings of the night i at lead-day j. We can fit observed LIFT_{ij} as a function of j with each data point weighted by the reciprocal of its estimated variance.

Fig. 9 demonstrates the relationship between lead-day and revenue lift with a Weighted Least Square (WLS) linear regression model[35] fitted. Even without observing the last month's data, we can see the linearly upward trend can project to the last month. We also see weighting the data points as the reciprocal of their variances is important when observed LIFT_{ij} has some large variance estimates between -120 to -100 lead days because less booking happened that early. Algorithm 1 details our entire fitting procedure.

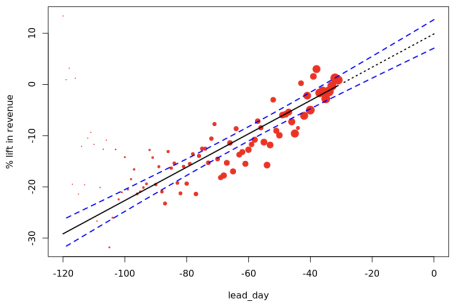

Figure 9: Revenue lift as a function of lead-day with the WLS regression line fitted. The treatment lowers the prices as the lead-day gets shorter, our revenue metric gets better over time, and this relationship can help to project the long-term revenue lift. The size of each data point is proportionate to the weight under WLS regression. Very few bookings happen between -120 to -100 lead day, thus, the variances of point-estimates are high and the weights are relatively small.

5 APPLY CORRECTION TO REAL EXPERIMENTS

For the real experiment introduced in Section 3.3, we use the nights from March to evaluate the performance of the remixed HTE estimator. We ran this experiment past March to get the final effect

Algorithm 1 Long-term effect prediction from short-term observations

function HTE($r^1_{ij}...r^{n_c}_{ij}, r^1_{ij}...r^{n_t}_{ij}$) ▷ This is short-term observation and we only have value of r_{ij} for long lead-day j=-120...-k where $k \gg 0$

$m^g_{ij} = \frac{1}{n_g}\left[\sum_{l=1}^{n_g} r^l_{ij}\right]$ ▷ $g=t,c$

$\text{LIFT}_{ij} = \frac{m^t_{ij} - m^c_{ij}}{m^c_{ij}}$ ▷ Calculate lift of revenue for night i booked j days ahead

$\hat{w}_{ij} = \frac{m^c_{ij}}{\sum_{j=-120}^{0} m^c_{ij}}$ ▷ $\hat{w}_{ij}$ is the average revenue taken in for night i at lead-day j across all the listings historically running the production strategy

Assume $\text{LIFT}_{ij} = f(j) + \epsilon_{ij}$, calculate $\hat{f}_{WLS}(j)$, the WLS linear regression estimator of f

$\widehat{\text{LIFT}}_{ij} = \hat{f}_{WLS}(j)$ ▷ j=-k+1...0

return: $\widehat{\text{LIFT}}^{hte} = \sum_i \sum_{j=-120}^{-k} \hat{w}_{ij} \times \text{LIFT}_{ij} + \sum_i \sum_{j=-k+1}^{0} \hat{w}_{ij} \times \widehat{\text{LIFT}}_{ij}$ ▷ The long-term effect estimate by remixed HTE

end function

estimate when all nights were realized. The question is: if we had stopped the experiment in Feb (short-term), can we still predict the March (long-term) outcome using the new remixed HTE estimator?

In order to test our estimator, we use long lead-day booking data of March nights made between the start of the experiment in December until February as the train set. Remixed HTE estimator is then applied on long lead-day revenue lifts to project out short lead-day revenue lifts to the final outcome. Our test set is the final experiment outcome at the end of March aggregated over all the bookings of March nights across all the lead-days. To estimate the confidence interval of the remixed HTE estimator, listing level bootstrap procedure[15] is applied. Algorithm 2 details the evaluation procedure.

Algorithm 2 Evaluating performance of remixed HTE estimator on a real experiment using bootstrap

for $b \leftarrow 1,B$ **do**

$(r'^1_{ij}...r'^{n_c}_{ij}, r'^1_{ij}...r'^{n_t}_{ij})$ ▷ Bootstrapped sample of the revenues of listings in the experiment, i being March nights, j=-120,...,0

$\text{LIFT}'^{short-term}$ ▷ Short-term experiment lift for j=-120,...,-k

$\text{LIFT}'^{long-term}$ ▷ Long-term experiment lift for j=-120,...,0

$\widehat{\text{LIFT}}'^{hte} = HTE(r'^1_{ij}...r'^{n_c}_{ij}, r'^1_{ij}...r'^{n_t}_{ij})$ ▷ HTE estimator of long-term lift using short-term booking data j=-120,...,-k

end for

return: Distribution of $\text{LIFT}'^{short-term}$, $\text{LIFT}'^{long-term}$, and $\widehat{\text{LIFT}}'^{hte}$

Fig. 10 demonstrates the performance of the estimator in relationship to the long-term and short-term experiment results. The naive short-term result deviates significantly from the long-term result. We could have discontinued running this experiment due to the significant estimated revenue loss using short-term booking data despite all other positive lifts in host satisfaction metrics. The remixed HTE estimator, in comparison, provides a confidence interval very close to the final effect estimate when all March nights were realized.

While we did not ship this experiment solely based off HTE result, the procedure gave us the assurance to keep this experiment running to convergence. Additionally, by running this experiment for a relatively long time, we have the data to validate our proposed approach. Tentatively, HTE could cut down the experiment run-time from four months to three months, and give us an initial read of the treatment impact as early as two months. Given the scale of Airbnb, being able to agilely ship or unship a pricing recommendation algorithm could translate into significant revenue growth for the hosts and platform.

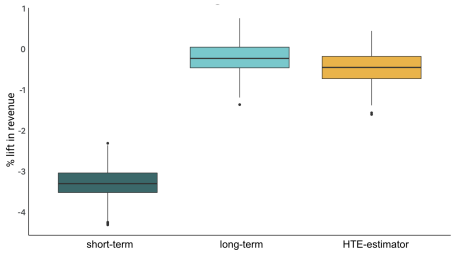

Figure 10: Remixed HTE experiment result from the short-term comes sufficiently close to the long-term result.

6 CONCLUSION

This paper raises the awareness of lead-day bias present in marketplace experimentation of expirable goods. A theoretical example was presented to demonstrate Mix Treatment Effect and Fill-rate Bias in play together to seriously distort the experiment result. Empirical evidence from a real pricing experiment supports the existence and highlights the magnitude of distortion that lead-day bias can cause.

To ameliorate lead-day bias, we propose a three-prong approach: (1) Limited roll-out of experiment: Instead of the traditional way of delivering treatment to all the nights of listings in the treatment group, we will only experiment on nights that are relatively far enough from the experiment start date and limit the impact analysis to those nights only. This will allow us to observe the performance of the pricing strategy for the entire booking cycle. (2) Smart-overlapping of pricing experiments with stack-up scheduling: By limiting treatment roll-out and experiment impact analysis to a specific set of nights, we can free up the other non-overlapping nights to other experiments. We can potentially triple the number of experiments that we can run in a calendar year and significantly accelerate our pace of innovation. (3) An HTE remixed estimator for long-term experiment impact: Instead of waiting for all the nights to be realized, we could adjust the short-term result to project out the long-term. Aside from significantly shortening the run-time of pricing experiments, the estimator can serve as the guardrail for us to avoid running a long experiment that could potentially hurt host revenues. Validation of our procedure on a long-running experiment shows considerable potential, and we will be investing heavily into validating and improving this solution. With these three workarounds, we are tripling the agility of our R&D process, making sure to deliver the best pricing insights to hosts in a timely manner.

Admittedly, there are other considerations to keep in mind in running marketplace experiments aside from the issues listed above. Seasonality could heavily impact the experiment result. Running the same experiment during the high vs. low season can point to different conclusions. Special events or the black swans can make the experiment result fail to generalize. The heterogeneity of treatment effect across listing segments or destinations is also worth considering in understanding the nuance of the treatment impact. Nevertheless, understanding and remedying lead-day bias in marketplace experiments is a milestone in our quest to innovate online measurement methodology.

ACKNOWLEDGMENTS

We would like to thank Ali Rauh for numerous fruitful feedbacks. We would also like to thank Minyong Lee for first hypothesizing the existence of lead-day bias.

REFERENCES

[1] Norman H Anderson and Alfred A Barrios. 1961. Primacy effects in personality impression formation. *The Journal of Abnormal and Social Psychology* 63, 2 (1961), 346.

[2] Susan Athey, Raj Chetty, and Guido Imbens. 2020. Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676* (2020).

[3] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2019. *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely.* Technical Report. National Bureau of Economic Research.

[4] Paolo Bartolomeo. 1997. The novelty effect in recovered hemineglect. *Cortex* 33, 2 (1997), 323–333.

[5] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Miruna Oprescu, and Vasilis Syrgkanis. 2021. Estimating the long-term effects of novel treatments. *Advances in Neural Information Processing Systems* 34 (2021), 2925–2935.

[6] Cameron A Belton and Robert Sugden. 2018. Attention and novelty: An experimental investigation of order effects in multiple valuation tasks. *Journal of Economic Psychology* 67 (2018), 103–115.

[7] Thomas Blake and Dominic Coey. 2014. Why marketplace experimentation is harder than it seems: The role of test-control interference. In *Proceedings of the fifteenth ACM conference on Economics and computation*. 567–582.

[8] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. 2022. Design and analysis of switchback experiments. *Management Science* (2022).

[9] Nanyu Chen, Min Liu, and Ya Xu. 2019. How A/B tests could go wrong: Automatic diagnosis of invalid online experiments. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 501–509.

[10] Lu Cheng, Ruocheng Guo, and Huan Liu. 2021. Long-term effect estimation with surrogate representation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 274–282.

[11] Alex Deng, Ulf Knoblich, and Jiannan Lu. 2018. Applying the Delta method in metric analytics: A practical guide with novel ideas. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 233–242.

[12] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 123–132.

[13] Pavel Dmitriev, Somit Gupta, Dong Woo Kim, and Garnet Vaz. 2017. A dirty dozen: twelve common metric interpretation pitfalls in online controlled experiments. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1427–1436.

[14] Dean Eckles, Brian Karrer, and Johan Ugander. 2017. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference* 5, 1 (2017).

[15] Bradley Efron. 1982. *The jackknife, the bootstrap and other resampling plans*. SIAM.

[16] Arne Feddersen, Wolfgang Maennig, Malte Borcherding, et al. 2006. The novelty effect of new soccer stadia: The case of Germany. *International Journal of Sport Finance* 1, 3 (2006), 174–188.

[17] Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter* 21, 1 (2019), 20–35.

[18] Henning Hohnhold, Deirdre O'Brien, and Diane Tang. 2015. Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1849–1858.

[19] David Holtz, Ruben Lobel, Inessa Liskovich, and Sinan Aral. 2020. Reducing interference bias in online marketplace pricing experiments. *arXiv preprint arXiv:2004.12489* (2020).

[20] Dennis R Howard and John L Crompton. 2003. An Empirical Review of the Stadium Novelty Effect. *Sport Marketing Quarterly* 12, 2 (2003).

[21] Edward E Jones, Leslie Rock, Kelly G Shaver, George R Goethals, and Lawrence M Ward. 1968. Pattern of performance and ability attribution: An unexpected primacy effect. *Journal of Personality and Social Psychology* 10, 4 (1968), 317.

[22] Raphael Lopez Kaufman, Jegar Pitchforth, and Lukas Vermeer. 2017. Democratizing online controlled experiments at Booking. com. *arXiv preprint arXiv:1710.08217* (2017).

[23] Ron Kohavi. 2015. Online controlled experiments: Lessons from running a/b/n tests for 12 years. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1–1.

[24] Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. 2012. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 786–794.

[25] Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. 2014. Seven rules of thumb for web site experimenters. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1857–1866.

[26] Ron Kohavi and Roger Longbotham. 2017. Online Controlled Experiments and A/B Testing. *Encyclopedia of machine learning and data mining* 7, 8 (2017), 922–929.

[27] Ron Kohavi, Diane Tang, and Ya Xu. 2020. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press.

[28] Hannah Li, Geng Zhao, Ramesh Johari, and Gabriel Y Weintraub. 2022. Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms. In *Proceedings of the ACM Web Conference 2022*. 182–192.

[29] Min Liu, Jialiang Mao, and Kang Kang. 2020. Trustworthy online marketplace experimentation with budget-split design. *arXiv preprint arXiv:2012.08724* (2020).

[30] Adity U Mutsuddi and Kay Connelly. 2012. Text messages for encouraging physical activity are they effective after the novelty effect wears off?. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. IEEE, 33–40.

[31] Cameron R Peterson and Wesley M DuCharme. 1967. A primacy effect in subjective probability revision. *Journal of Experimental Psychology* 73, 1 (1967), 61.

[32] Jordan Poppenk, Stefan Köhler, and Morris Moscovitch. 2010. Revisiting the novelty effect: when familiarity, not novelty, enhances memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36, 5 (2010), 1321.

[33] Soheil Sadeghi, Somit Gupta, Stefan Gramatovici, Jiannan Lu, Hao Ai, and Ruhan Zhang. 2022. Novelty and primacy: a long-term estimator for online experiments. *Technometrics* 64, 4 (2022), 524–534.

[34] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. 2017. Detecting network effects: Randomizing over randomized experiments. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1027–1035.

[35] Tilo Strutz. 2011. *Data fitting and uncertainty: A practical introduction to weighted least squares and beyond*. Springer.

[36] Lydia Tan and Geoff Ward. 2000. A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26, 6 (2000), 1589.

[37] Edward L Thorndike. 1898. Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements* 2, 4 (1898), i.

[38] Patrick FA Van Erkel and Peter Thijssen. 2016. The first one wins: Distilling the primacy effect. *Electoral Studies* 44 (2016), 245–254.

[39] Tyler J VanderWeele. 2013. Surrogate measures and consistent surrogates. *Biometrics* 69, 3 (2013), 561–565.

[40] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[41] Huizhi Xie and Juliette Aurisset. 2016. Improving the sensitivity of online controlled experiments: Case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 645–654.

[42] Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2227–2236.

[43] Peng Ye, Julian Qian, Jieying Chen, Chen-hung Wu, Yitong Zhou, Spencer De Mars, Frank Yang, and Li Zhang. 2018. Customized regression model for airbnb dynamic pricing. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 932–940.