23

Measuring Long-Term Treatment Effects

> We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run
> – *Roy Amara*

***Why you care**: Sometimes the effect that you care to measure can take months or even years to accumulate – a long-term effect. In an online world where products and services are developed quickly and iteratively in an agile fashion, trying to measure a long-term effect is challenging. While an active area of research, understanding the key challenges and current methodology is useful if you are tackling a problem of this nature.*

What Are Long-Term Effects?

In most scenarios discussed in this book, we recommend running experiments for one to two weeks. The Treatment effect measured in this short timeframe is called the *short-term* effect. For most experiments, understanding this short-term effect is all we need, as it is stable and generalizes to the *long-term* Treatment effect, which is usually what we care about. However, there are scenarios where the long-term effect is different from the short-term effect. For example, raising prices is likely to increase short-term revenue but reduce long-term revenue as users abandon the product or service. Showing poor search results on a search engine will cause users to search again (Kohavi et al. 2012); the query share increases in the short-term but decreases in the long-term as users switch to a better search engine. Similarly, showing more ads – including more low-quality ads – can increase ad clicks and revenue in the short-term but decreases revenue via decreased ad clicks, and even searches, in the long-term (Hohnhold, O'Brien and Tang 2015, Dmitriev, Frasca, et al. 2016).

The long-term effect is defined as the asymptotic effect of a Treatment, which, in theory, can be years out. Practically, it is common to consider long-term to be 3+ months, or based on the number of exposures (e.g., the Treatment effect for users exposed to the new feature at least 10 times).

We explicitly exclude from discussion changes that have a short life span. For example, you may run an experiment on news headlines picked by editors that have a life span of only a few hours. However, the question of whether headlines should be "catchy" or "funny" is a good long-term hypothesis, as an initial increase in short-term engagement may also be associated with long-term increased abandonment. Except when you are specifically running experiments on such short life-span changes, when testing a new Treatment, you would really like to know how it would perform in the long term.

In this chapter, we cover the reasons that long-term effects can be different from short-term effects and discuss measurement methods. We only focus on scenarios where short-term and long-term Treatment effects differ. We are not considering other important differences between short-term and long-term, such as sample size difference, which may cause the *estimated* Treatment effects and variance to differ.

One key challenge in determining the OEC (see Chapter 7) is that it must be measurable in the short term but believed to causally impact long-term objectives. Measuring long-term effects discussed in this chapter can provide insights to improve and devise short-term metrics that impact the long-term goals.

Reasons the Treatment Effect May Differ between Short-Term and Long-Term

There are several reasons why short-term and long-term Treatment effects may differ. We have discussed some in the context of trustworthiness in Chapter 3.

- **User-learned effects**. As users learn and adapt to a change, their behavior changes. For example, product crashes are a terrible user experience that may not turn users away with the first occurrence. However, if crashes are frequent, users learn and may decide to leave the product. Users may adjust the rate they click on ads if they realize the ads' quality is poor. The behavior change may also be due to discoverability, maybe a new feature that may take time for users to notice, but once they discover its usefulness, they engage heavily. Users may also need time to adapt to a new feature because they are primed in the old feature, or they explore a new change

more when it is first introduced (see Chapter 3). In such cases, a long-term effect may differ from a short-term effect because users eventually reach an equilibrium point (Huang, Reiley and Raibov 2018, Hohnhold, O'Brien and Tang 2015, Chen, Liu and Xu 2019, Kohavi, Longbotham et al. 2009).

- **Network effects**. When users see friends using the Live Video feature on a communication app such as Facebook Messenger, WhatsApp, or Skype, it is more likely that they will use it too. User behavior tends to be influenced by people in their network though it may take a while for a feature to reach its full effect as it propagates through their network (see Chapter 22, which discusses interference in marketplaces with limited or shared resources, focusing on biased estimation in the *short-term* due to leakage between variants). The limited resources introduce additional challenges as we measure long-term impact. For example, in two-sided marketplaces, such as Airbnb, eBay, and Uber, a new feature can be very effective at driving demand for an item, such as a house to rent, computer keyboard, or ride, but the supply may take longer to catch up. As a result, the impact on revenue may take longer to realize as supply is unavailable. Similar examples exist for other areas, such as hiring marketplaces (job seekers and jobs), ad marketplaces (advertisers and publishers), recommendation systems for content (news feeds), or connections (LinkedIn's People You May Know). Because there are a limited number of people one person knows ("supply"), a new algorithm may perform better at the beginning but may reach a lower equilibrium long term because of supply constraints (an analogous effect can be seen in recommendation algorithms more generally, where a new algorithm may perform better initially due to diversity, or simply showing new recommendations).
- **Delayed experience and measurement.** There can be a time gap before a user experiences the entirety of the Treatment effect. For example, for companies like Airbnb and Booking.com, there can be months between a user's online experience and when the user physically arrives at the destination. The metrics that matter, such as user retention, can be affected by the user's delayed offline experience. Another example is annual contracts: Users who sign up have a decision point when the year ends and their cumulative experience over that year determines whether they renew.
- **Ecosystem change.** Many things in your ecosystem change over time and can impact how users react to the Treatment, including:
  - **Launching other new features**. For example, if more teams embed the Live Video feature in their product, Live Video becomes more valuable.
  - **Seasonality**. For example, experiments on gift cards that perform well during the Christmas season may not have the same performance during the non-holiday season due to users having different purchasing intent.

- **Competitive landscape**. For example, if your competition launches the same feature, the value of the feature may decline.
- **Government policies**. For example, the European Union General Data Protection Regulation (GDPR) changes how users control their online data, and hence what data you can use for online ad targeting (European Commission 2016, Basin, Debois and Hildebrandt 2018, Google, Helping advertisers comply with the GDPR 2019).
- **Concept drift**. The performance of machine learning models trained on data that is not refreshed may degrade over time as distributions change.
- **Software rot**. After features ship, unless they are maintained, they tend to degrade with respect to the environment around them. This can be caused, for example, by system assumptions made by code that becomes invalid over time.

Why Measure Long-Term Effects?

While the long-term effect can certainly differ from short-term effect for various reasons, not all such differences are worth measuring. What you want to achieve with the long-term effect plays a critical role in determining what you should measure and how you should measure it. We summarize the top reasons.

- **Attribution**. Companies with strong data-driven culture use experiment results to track team goals and performance, potentially incorporating experiment gains into long-term financial forecasting. In these scenarios, proper measurement and attribution of the long-term impact of an experiment is needed. What would the world look like in the long term with vs. without introducing the new feature now? This type of attribution is challenging because we need to consider both endogenous reasons such as user-learned effects, and exogenous reasons such as competitive landscape changes. In practice, because future product changes are usually built on top of past launches, it may be hard to attribute such compounding impacts.
- **Institutional learning**. What is the difference between short term and long term? If the difference is sizable, what is causing it? If there is a strong novelty effect, this may indicate a suboptimal user experience. For example, if it takes a user too long to discover a new feature they like, you may expedite uptake by using in-product education. On the other hand, if many users are attracted to the new feature but only try it once, it may indicate low quality or click-bait. Learning about the difference can offer insights into an improved subsequent iteration.

- **Generalization**. In many cases, we measure the long-term effect on some experiments so we can extrapolate to other experiments. How much long-term impact does a similar change have? Can we derive a general principle for certain product areas (e.g., search ads in Hohnhold et. al. (2015)? Can we create a short-term metric that is predictive of long term (see the last section of this chapter)? If we can generalize or predict the long-term effect, we can take that generalization into account in the decision-making process. For this purpose, you may want to isolate the long-term impact from exogenous factors, especially big shocks that are unlikely to repeat over time.

Long-Running Experiments

The simplest and most popular approach for measuring long-term effects is to keep an experiment running for a long time. You can measure the Treatment effect at the beginning of the experiment (in the first week) and at the end of the experiment (in the last week). Note that this analysis approach differs from a typical experiment analysis that would measure the average effect over the entire Treatment period. The first percent-delta measurement $p\Delta_1$ is considered the short-term effect and the last measurement $p\Delta_T$ is the long-term effect as shown in Figure 23.1.

While this is a viable solution, there are several challenges and limitations in this type of long-term experiment design. We focus on a few that are relevant

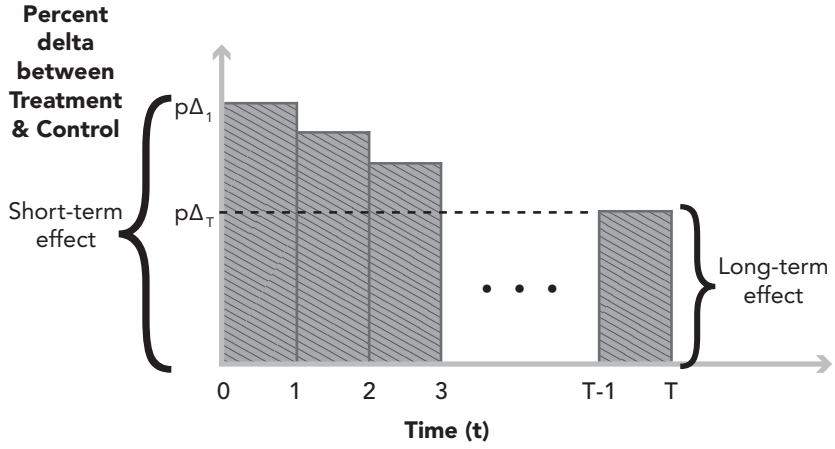

Figure 23.1 Measuring long-term effect based on a long-running experiment

to measuring long-term effects, organized around the purpose of attribution and institutional learning.

- **For attribution**: The measurement from the last week of the long-running experiment ($p\Delta_T$) may not represent the true long-term Treatment effect for the following reasons:
  - Treatment effect dilution.
  - The user may use multiple devices or entry points (e.g., web and app), while the experiment is only capturing a subset. The longer the experiment runs, the more likely a user will have used multiple devices during the experiment period. For users who visit during the last week, only a fraction of their experience during the entire time period T is actually in Treatment. Therefore, if users are learning, what is measured in $p\Delta_T$ is not the long-term impact of what users learned after being exposed to Treatment for time T, but a diluted version. Note that this dilution may not matter for all features, but rather the subset where the dosage matters.
  - If you randomize the experiment units based on cookies, cookies can churn due to user behavior or get clobbered due to browser issues (Dmitriev et al. 2016). A user who was in Treatment could get randomized into Control with a new cookie. As in the two previous bullet points, the longer the experiment runs, the more likely that a user will have experienced both Treatment and Control.
  - If network effects are present, unless you have perfect isolation between the variants, the Treatment effect can "leak" from Treatment to Control (see Chapter 22). The longer the experiment runs, it is likely that the effect will cascade more broadly through the network, creating larger leakage.
- **Survivorship bias.** Not all users at the beginning of the experiment will survive to the end of the experiment. If the survival rate is different between Treatment and Control, $p\Delta_T$ would suffer from *survivorship bias*, which should also trigger an SRM alert (see Chapters 3 and 21). For example, if those Treatment users who dislike the new feature end up abandoning over time, $p\Delta_T$ would only capture a biased view of those who remain (and the new users admitted to the experiment). Similar bias can also exist if the Treatment introduces a bug or side-effect that causes a different cookie churn rate.
- **Interaction with other new features**. There can be many other features launched while the long-term experiment is running, and they may interact with the specific feature being tested. These new features can erode the wins of the experiment over time. For example, a first experiment that sends push

notifications to users can be hugely effective at driving sessions, but as other teams start sending notifications, the effect of the first notification diminishes.

- **For measuring a time-extrapolated effect**: Without further study – including more experiments – we need to be cautious to not interpret the difference between $p\Delta_0$ and $p\Delta_T$ as a meaningful difference caused by the Treatment itself. Besides the attribution challenges discussed above that complicate the interpretation of $p\Delta_T$ itself, the difference may be purely due to exogenous factors, such as seasonality. In general, if the underlying population or external environment has changed between the two time periods, we can no longer directly compare short-term and long-term experiment results.

Of course, challenges around attribution and measuring time-extrapolated effects also make it hard to generalize the results from specific long-running experiments to more extensible principles and techniques. There are also challenges around how to know that the long-term result is stabilized and when to stop the experiment. The next section explores experiment design and analysis methodologies that partially address these challenges.

Alternative Methods for Long-Running Experiments

Different methods have been proposed to improve measurements from long-running experiments (Hohnhold, O'Brien and Tang 2015, Dmitriev, Frasca, et al. 2016). Each method discussed in this section offers some improvements, but none fully address the limitations under all scenarios. We highly recommend that you always evaluate whether these limitations apply, and if so, how much they impact your results or your interpretation of the results.

Method #1: Cohort Analysis

You can construct a stable cohort of users before starting the experiment and only analyze the short-term and long-term effects on this cohort of users. One method is to select the cohort based on a stable ID, for example, logged-in user IDs. This method can be effective at addressing dilution and survivorship bias, especially if the cohort can be tracked and measured in a stable way. There are two important considerations to keep in mind:

- You need to evaluate how stable the cohort is, as it is crucial for the effectiveness of the method. For example, if the ID is based on cookies

and when cookie churn rate is high, this method does not work well for correcting bias (Dmitriev et al. 2016).

- If the cohort is not representative of the overall population, there may be external validity concerns because the analysis results may not be generalizable to the full population. For example, analyzing logged-in users only may introduce bias because they differ from non-logged-in users. You can use additional methods to improve the generalizability, such as a weighting adjustment based on stratification (Park, Gelman and Bafumi 2004, Gelman 1997, Lax and Phillips 2009). In this approach, you first stratify users into subgroups (e.g., based on pre-experiment high/medium/low engagement levels), and then compute a weighted average of the Treatment effects from each subgroup, with the weights reflecting the population distribution. This approach has similar limitations as observational studies discussed extensively in Chapter 11.

Method #2: Post-Period Analysis

In this method, you turn off the experiment after it has been running for a while (time T) and then measure the difference between the users in Treatment and those in Control during time T and $T+1$, as shown in Figure 23.2. In the event where you cannot ramp down the new Treatment due to user experience concerns, you can still apply this method by "ramping up" the Treatment for all users. A key aspect of this method is that during the measurement period, users in the Treatment and Control groups are both exposed to the exact same

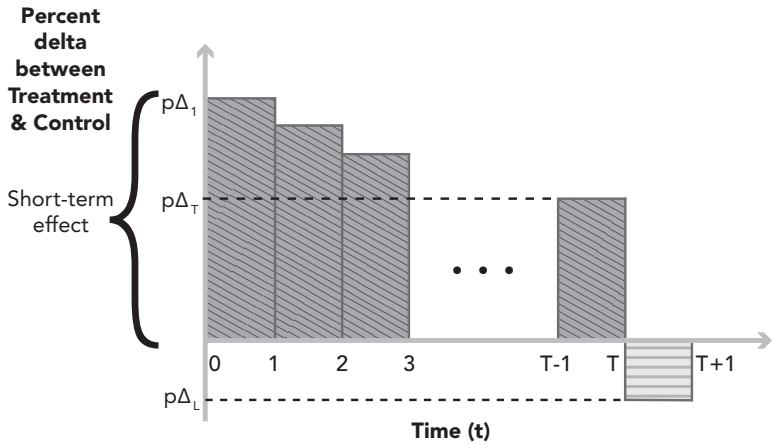

Figure 23.2 Measuring long-term effect based on post-period A/A measurement

features. The difference between the groups, however, is that in the first case, the Treatment group was exposed to a set of features that the Control group was not exposed to, or in the second "ramping up" case, the Treatment group was exposed to the features for a longer time than the Control group.

Hohnhold et al. (2015) calls the effect measured during the post-period the *learning effect*. To properly interpret it, you need to understand the specific change tested in the experiment. There are two types of learned effect:

1. **User-learned effect**. Users have learned and adapted to the change over time. Hohnhold et al. (2015) studies the impact of increasing ad-load on users' ad clicking behavior. In their case study, user learning is considered the key reason behind the post-period effect.
2. **System-learned effect**. The system may have "remembered" information from the Treatment period. For example, the Treatment may encourage more users to update their profile and this updated information stays in the system even after the experiment ends. Or, if more Treatment users are annoyed by emails and opt out during the experiment, they will not receive emails during the post-period. Another common example is personalization through machine learning models, such as models that show more ads to users who click more on ads. After a Treatment that causes users to click more on ads, the system that uses a sufficiently long time period for personalization may learn about the user and thus show them more ads even after they are back to experiencing the Control Treatment.

Given enough experiments, the method can estimate a learned effect based on the system parameters and subsequently extrapolate from new short-term experiments to estimate the anticipated long-term effect (Gupta et al. 2019). This extrapolation is reasonable when the system-learned effects are zero, that is, in the A/A post-period, both Treatment and Control users are exposed to the exact same set of features. Examples of where this system-learned effect is non-zero might include permanent user state changes, such as more time-persistent personalization, opt-outs, unsubscribes, hitting impression limits, and so on.

That said, this approach is effective at isolating impact from exogenous factors that change over time and from potential interactions with other newly launched features. Because the learned effect is measured separately, it offers more insights on why the effects are different short term vs. long term. This method suffers from potential dilution and survivorship bias (Dmitriev et al. 2016). However, because the learned effect is measured separately in the post-period, you could attempt to apply an adjustment to the learned effect to account for dilution or by combining with the cohort analysis method discussed earlier.

Method #3: Time-Staggered Treatments

The methods discussed so far simply require experimenters to wait "enough" time before taking the long-term measurement. But how long is "long enough?" A poor man's approach is to observe the Treatment effect trend line and decide that enough time has passed when the curve stabilizes. This does not work well in practice because Treatment effect is rarely stable over time. With big events or even day-of-week effect, the volatility over time tends to overwhelm the long-term trend.

To determine the measurement time, you can have two versions of the same Treatment running with staggered start times. One version (T_0) starts at time $t=0$, while the other (T_1) starts at time $t=1$. At any given time, $t>1$, you can measure the difference between the two versions of Treatment. Note that at time t, T_0 and T_1 are effectively an A/A test with the only difference being the duration their users are exposed to Treatment. We can conduct a two-sample t-test to check whether the difference between $T_1(t)$ and $T_0(t)$ is statistically significant, and conclude that the two Treatments have converged if the difference is small, as shown in Figure 23.3. Note that it is important to determine the practically significant delta and ensure that the comparison has enough statistical power to detect it. At this point, we can apply the post-period

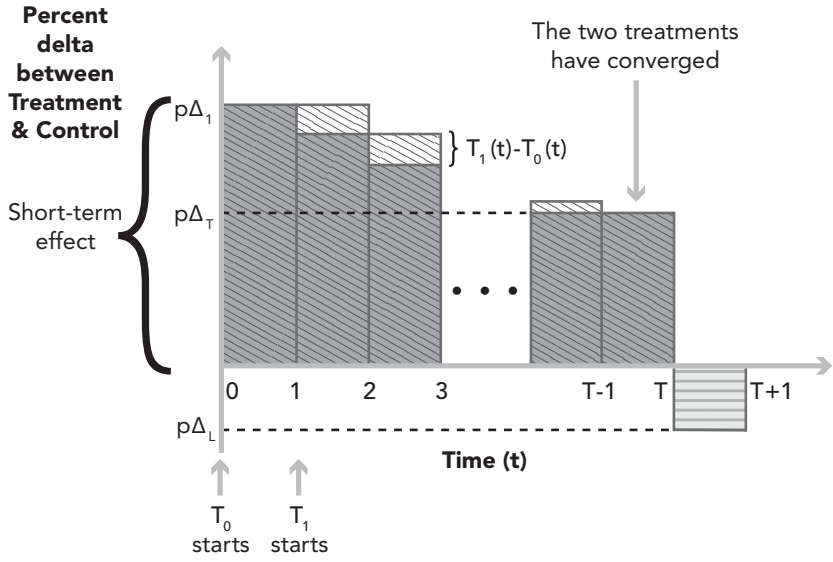

Figure 23.3 Measuring long-term effect after we observe the two time-staggered Treatments have converged

method after time t to measure the long-term effect (Gupta, Kohavi et al. 2019). While testing the difference between the two Treatments, it may be more important to control for a lower type II error rate than the typical 20%, even at the cost of increasing the Type I error rate to be higher than 5%.

This method assumes that the difference between the two Treatments grows smaller over time. In other words, $T_1(t) - T_0(t)$ is a decreasing function of t. While this is a plausible assumption, in practice, you also need to ensure that there is enough time gap between the two staggered Treatments. If the learned effect takes some time to manifest, and the two Treatments start right after one another, there may not be enough time for the two Treatments to have a difference at the start of T_1.

Method #4: Holdback and Reverse Experiment

Long-term experiments may not be feasible if there is time pressure to launch a Treatment to all users. Control groups can be expensive: they have an opportunity cost as they don't receive the Treatment (Varian 2007). An alternative is to conduct a *holdback*: keeping 10% of users in Control for several weeks (or months) after launching the Treatment to 90% users (Xu, Duan and Huang 2018). Holdback experiments are a typical type of long-running experiment. Because they have a small Control variant, they tend to have less power than may be optimal. It is important to make sure that the reduced sensitivity does not impact what you want to learn from the holdout. See more discussion in Chapter 15.

There is an alternative version called *reverse* experiments. In a reverse experiment, we ramp 10% of users back into the Control several weeks (or months) after launching the Treatment to 100% of users. The benefit of this approach is that everyone has received the Treatment for a while. If the Treatment introduces a new feature where network effect plays a role in user adoption, or if supply is constrained in the marketplace, the reverse experiment allows the network or the marketplace time to reach the new equilibrium. The disadvantage is that if the Treatment may introduce a visible change, ramping the users back into the Control may confuse them.