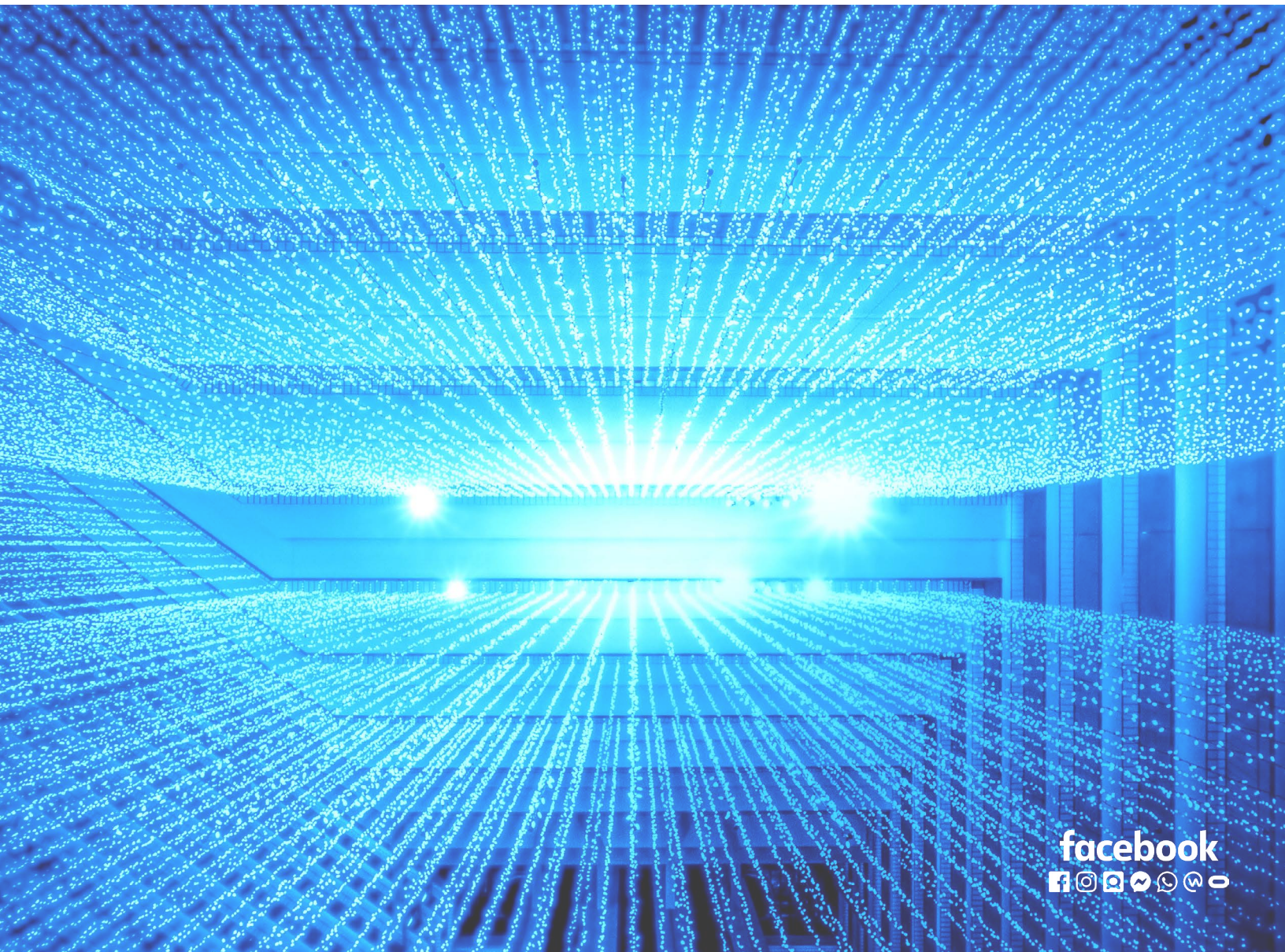
MARCH 2019

Prepare for the Unexpected

A Guide to Testing and Learning with Incrementality Measurement



Contents

Contributors

Alok Gupta, Lyft
Director of Data Science

Matthew Gerrie, Booking.com
Senior Director of Marketing Science and Communication

Stephan McBride, Netflix
Director of Science and Analytics, Marketing and Economics

Tony Flanery-Rye, eBay
Senior Director of Growth Analytics

Dan Johns, Facebook
Product Marketing Manager

Jesse Goranson, Facebook
Director of Client Measurement

Maggie Burke, Facebook
Client Council Lead

Sophia Lin, Facebook
Project Manager

Introduction

You're ready to start testing different marketing strategies, but are you prepared for all the ways these tests might not go as planned?

In "Measure Marketing Effectiveness: A Guide to Implementing Incrementality," we showed how to measure a marketing strategy's true value. With incrementality measurement, marketers can gauge the true value of strategy by isolating it from other strategies, business factors and variables.

According to leading marketers, the most effective way to gauge the value of a strategy is through experimentation. By nature, experiments are subject to external factors and internal assumptions. Experiments can also be unpredictable, and at times some even go wrong. But even when they do, these tests can still prove useful and provide learnings.

In partnership with top measurement experts, we created this report to provide a practical guidebook for what can go wrong during marketing experimentation and ideas for what to do when things do. Marketers can still fully reap the benefits of incrementality measurement, even with potential testing challenges along the way.

"You should always assume and expect that a lot of things will go wrong," said **Stephan McBride, Director of Science and Analytics, Marketing and Economics, Netflix.** "Really invest in preparing for that because success in incrementality means that experimental evidence will be given tremendous weight in your business."

The good news is the things that can go wrong are often predictable, meaning marketers can plan for them. Even better news: the typical challenges that arise from experiments can be instructive—maybe even transformative—for marketers. Indeed, the most successful brands over the last few years have a philosophy in common—they have embraced a fearless, test-and-learn culture and have found that is the most effective way for a business to grow and compete.

In preparing for what to do when things go wrong, marketers can begin this journey ready to address any challenges, armed with the knowledge of which experiments are valuable regardless of what might happen.

"I don't know if there is a time where I ran a larger experiment where something didn't go wrong," said **Tony Flanery-Rye, Senior Director of Growth Analytics, eBay.** "But it's okay. You need to be adaptable to these events."

By making testing and learning a core part of your business, you can help your organization achieve marketing excellence.

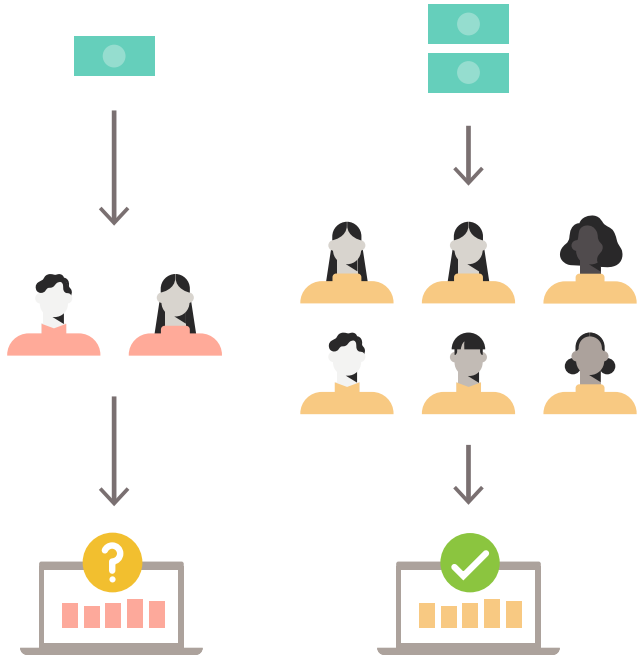
PROBLEM 1

Your Test Doesn't Have Enough Statistical Power

When conducting a test to measure incrementality, having the right amount of data is crucial.

A small sample size can skew results, while too large of a data sample can be costly. Business judgment must balance the need for learning with the cost of larger and longer campaigns. Ultimately, your test is only as good as its statistical power.

For example, when running experiments, make sure that your test is set up with enough precision to measure what you're trying to detect. It will take more data to reveal small differences and less data to see large differences. Even if you've planned a test up front to provide enough data, planning based on historical data might not hold up during a test in the real world. In fact, not every test will have enough statistical power to detect a change in performance.



How can this happen?

Not having statistical power can be caused by a number of factors:

There simply isn't enough distinction between your test and control group. Sometimes there simply isn't an effect, or difference, between the treatments you're testing. No matter how much data you gather, you won't confidently be able to detect a change.

The results you see are different than what you expected based on past experiments. The effect of the treatment could be lower than you thought based on historical testing. When this happens, you might have expected to need less data and, as a result, your test didn't collect enough samples to detect the true effect. This means there might be an effect, but the experiment cannot detect it.

You weren't able to collect as much data as you wanted because of external factors. External events can prevent your campaign from delivering as expected or influence your KPI (such as conversion rates, sales, app installs or store revisits). This can change the amount of data you collect and lead to tests that are underpowered.

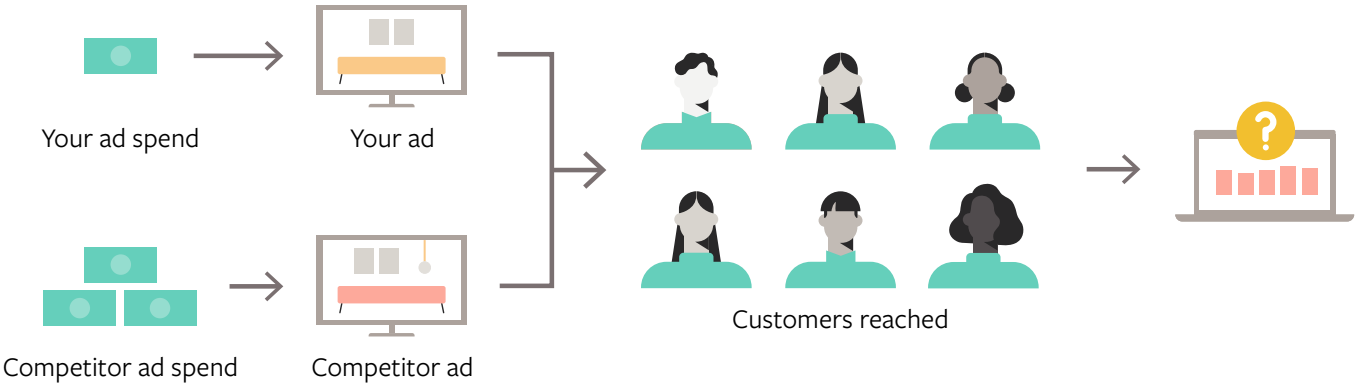
You don't have enough data for the particular experiment being conducted. If you're testing against a defined audience, such as a customer list or people who use a certain feature, you might never be able to reach enough people to detect an effect.

The outcomes you're testing just don't occur that often. Some outcomes happen infrequently and make it hard to collect enough data to get statistically significant results.

A long purchase cycle makes it hard to collect enough data. For outcomes that have long consideration cycles, you may not have run an experiment for a long enough time to truly see an impact. Issues such as cookie contamination and other forms of error can also dilute your effect, thereby reducing statistical power.

What could this look like?

A budding direct-to-consumer mattress seller has built a customer database primarily through social media advertising and now is ready to start testing TV ad spend for the first time to gauge its impact on sales. But just as the seller is ready to kick off a new TV flight in some cities within a matched market test, another mattress upstart kicks into gear on TV, spending at a much higher level and thus reaching a lot more consumers. As results begin to come in, the seller realizes the test doesn't have enough statistical power to detect a change in performance. Did the TV ads work? Or were they drowned out by the competitor's campaign?



What could you do?

In the above example, a lack of statistical power doesn't necessarily mean a test has to be discarded. In fact, it's possible that the seller may still be able to use the results. One way to determine that is through a power calculation, which is calculating the probability that the study will be able to detect a lift if there is actually a lift. This is a vital indicator of whether there will be enough data to report reliable results.

The intent of a power calculation is to detect, with some level of certainty, an effect over a certain size (*I can detect effects bigger than x%*). Simply not detecting an effect doesn't mean there wasn't one. If you didn't collect enough data, you effectively change the certainty threshold you can detect. If there isn't an effect over a certain size that you were expecting, that doesn't mean there wasn't an effect size smaller than that.

Besides gauging whether you have enough data to determine the impact, such a test can still be directionally useful. For instance, if you need to see an effect at a specific percentage to make the strategy profitable, and you know your effect doesn't meet that, you can still make a data-driven decision.

Or, if you're comparing treatments, you can still use the results to determine the confidence in your findings. While the power of the study may not be as high as you'd like (which should be greater than 80%, generally speaking), you might still be able to make a decision based on this information. When choosing between two or more options for strategy execution, you still want to pursue the one that's most likely to provide more incrementality, all things else equal.

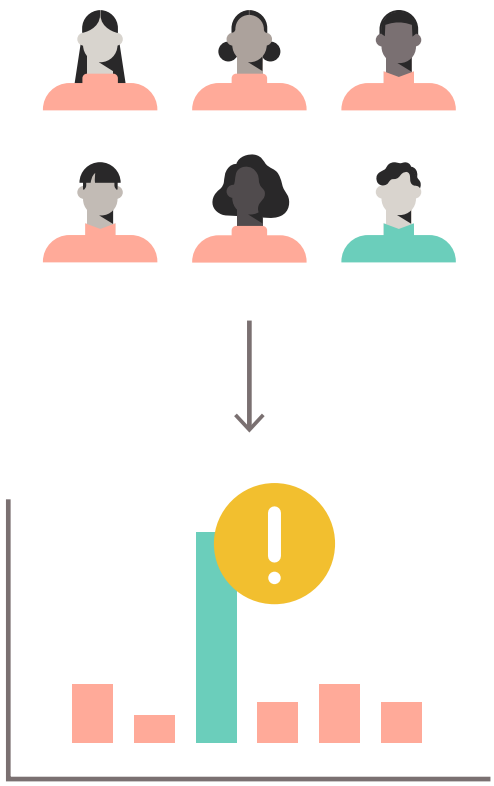
PROBLEM 2

One of Your Test Groups Has Some Outliers

Incrementality inherently depends on comparing two or more groups.

Randomized assignment within a marketing experiment helps ensure that the groups are statistically similar. However, outliers can make comparisons problematic. Outliers can be defined as test participants who exhibit results that are well above or below the norm.

Outliers can make two groups being tested dissimilar. It's true that outliers exist in all data sets, but their presence in a test can dramatically change how you evaluate a treatment. For smaller tests, the presence of an extremely large customer in one of the treatment groups can sway the results.



How can this happen?

As a first step, it's important to determine why these outliers could exist for your business:

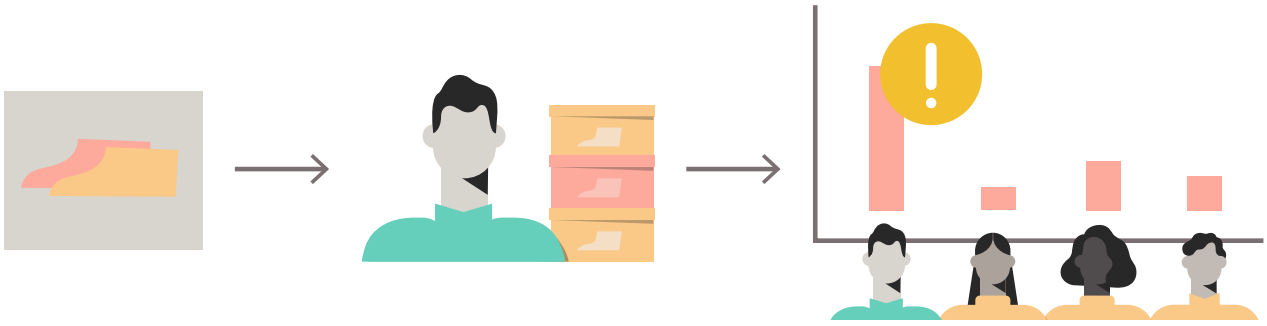
Your business is driven by outliers.
Many businesses that operate on a freemium model, like gaming and some SaaS companies, are built around infrequent, large purchasers. Gaming companies often refer to these customers as whales, as they account for a much larger portion of revenue than the average player. SaaS companies, on the other hand, might sell software that is available to individuals at a monthly fee, but a corporation may buy access to this software for hundreds of employees, spending several hundred thousands of dollars in a single month.

You have a mix of average spending consumers and big spending businesses.
Your customer base includes a mix of consumers with infrequent, low-value transactions and business customers with frequent or high-value transactions, and there is significant assignment imbalance among these sets.

Sales numbers were impacted by random events.
A random event during the test causes people to act differently than they normally would. For example, a travel agency might have a normal business customer hosting a sales conference and booking flights for all 7,000 attendees.

What could this look like?

A small retail business selling custom footwear for runners is pushing a holiday sale, hoping to acquire a slew of new customers while measuring the effectiveness of the sale campaign. Upon seeing the ads, a CEO decides to order custom sneakers for his entire company as a holiday gift. On first glance, it seems as if the campaign was a wild success. However, upon further investigation, the small business owner realizes that a single purchaser accounts for the majority of sales.



What could you do?

Be on the lookout from the start. Make sure your analytics or data science team investigates for the presence of outliers. While you can prepare yourself for outliers ahead of time during statistical power calculations, we also recommend developing a strategy for identifying and adjusting for outliers after the fact.

Winsorization is one common statistical tactic used to adjust for outliers by imputing a more common value for the outlier. It is a process of compartmentalizing outliers by isolating their data to a specified percentile.

For example, following a marketing test that was skewed by outlier data, a company might elect to use the median sales numbers in place of the top 0.01% of customers, rather than their actual sales, since they are so large they're likely to throw off the data.

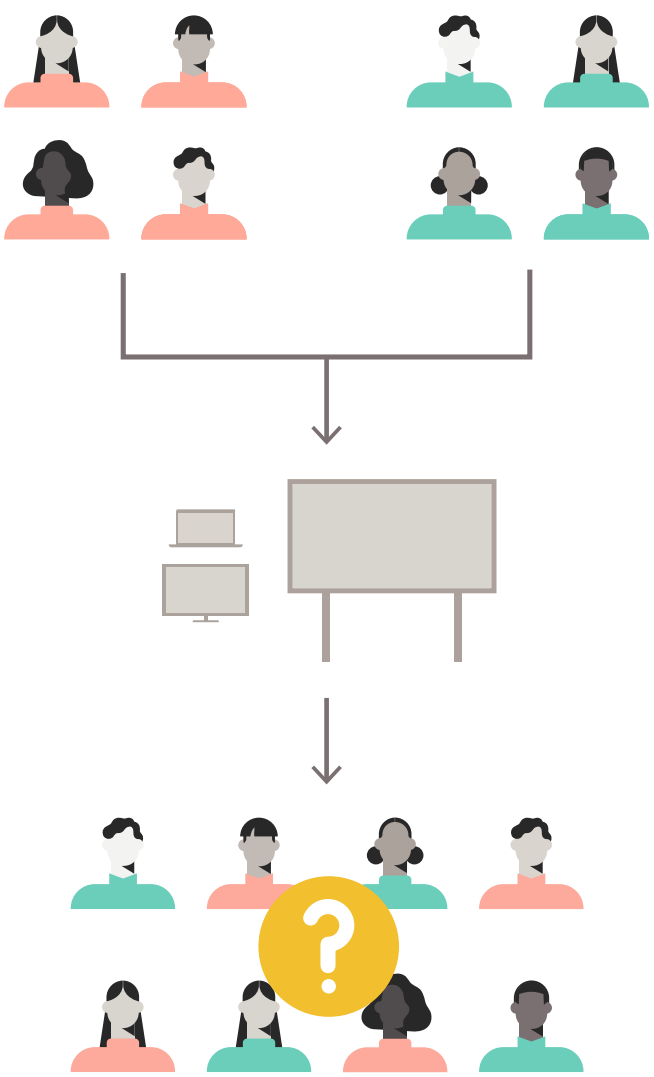
This works well when the outliers are not a regular part of your business or aren't caused by errors. If the source of the outlier is not caused by a legitimate business reason, such as an error, you might want to use a different approach and remove the outlier entirely through exclusion.

PROBLEM 3

Your Variables Are Not Isolated

Isolating variables allows you to accurately understand what causes the change you observe in an experiment.

But there are variables you can control, variables you can't and those you can't foresee. Despite your best efforts, people might be exposed to other treatments during the course of a test, resulting in your variables not being isolated.



How can this happen?

There are numerous reasons why a particular test's variables may not be properly isolated:

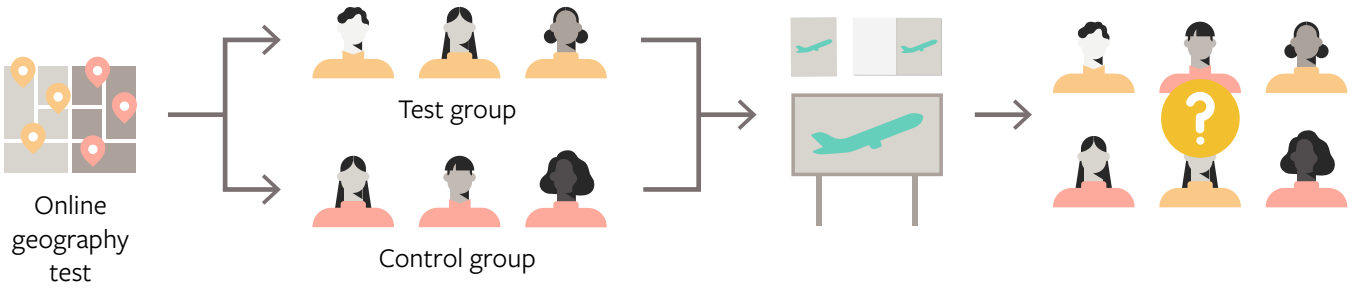
Not everybody is aligned during execution, leading to errors.
When running a media campaign, the various teams involved in execution may not have kept all other variables the same across treatments. They might have adjusted bids, budgets, creatives or other executional details on one treatment group and not another.

You don't control all of your company's media plans.
People in your treatment group might have been exposed to more than one change because a partner team is running unexpected media or outreach programs unevenly across your treatment groups. It's worth noting that you shouldn't stop other channels while running a test. In fact, you want to run a test under the same conditions you'd roll out the results—thus, you want to keep running CRM, performance marketing and other activities in the same way you'd run them under usual conditions. The key is to make sure those programs are run evenly across test and control groups.

Members of your test group affect your control group, limiting how controlled it is.
If an action by a user in one treatment affects the behavior of a user in another treatment, then the control will not truly reflect the absence of an intervention. For example, if a customer in the treatment group purchases a book that is in limited supply, it may affect the individuals in a control group who would have also purchased the same book but are now unable to due to the lack of supply available.

What could this look like?

A travel company is running an online video campaign to make people aware of its summer sale. To test the effectiveness of this campaign, the company's marketer decides to conduct an online geography test which removes people in specific zip codes or DMAs (Designated Market Areas). Unaware of the test that is running, the company's outdoor media manager puts up out-of-home advertising in select cities. This causes people in these hold-out groups to be exposed to the summer sale, even though they were intended to be the control group. The variables are no longer isolated, and it is difficult to understand the effect of the treatment.



What could you do?

While variables might not have been entirely isolated, this does not mean the results are unusable. If the differences were small, like minor executional changes, you may be able to safely ignore this error if the effect you're observing is big enough.

If the errors were large, you can still use the results, but you might have to either change the interpretation of the test or model out any bias since you cannot confidently isolate the cause of the change to what you originally designed in the experiment.

In the event that you want to retest, you will now have more information about expected effect sizes or potential execution errors.

PROBLEM 4

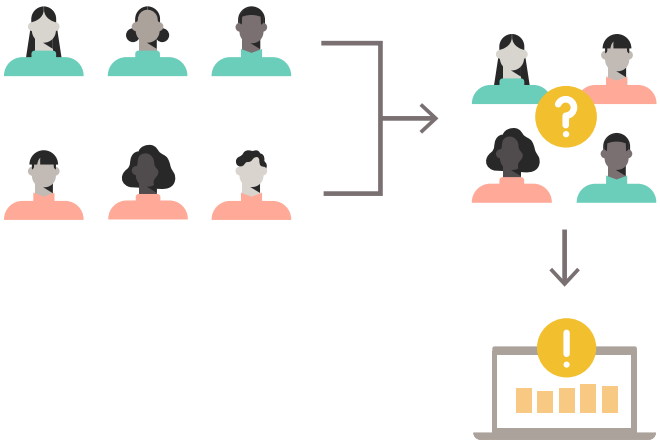
People in Your Test and Control Groups Cross Paths

A key component of any experiment is making sure members of each test audience stay in the treatment groups they are assigned to throughout the length of the test and across whichever devices and platforms you're measuring.

That way, people who aren't supposed to experience a treatment (such as a creative test) don't see it, and people who are supposed to see a treatment have the opportunity to see it at your intended cadence.

But the real world doesn't always work that way. People are unpredictable. For example, children might be watching videos on their parents' phones and see ads aimed at their moms or dads, urging them to buy a new minivan. Or a fast food chain could be promoting a new breakfast sandwich in certain markets by testing a new local media strategy, yet some people from other parts of the country may be traveling and encounter these test ads, even though this product is not available where they live.

It's possible that during a test, some people might be exposed to multiple treatments. Alternatively, some people in your control group could also be exposed to a treatment. Finally, some people may become unexposed by being moved to the control group, weakening the signal.



How can this happen?

There are several different ways test and control groups can inadvertently end up crossing paths:

The way you are identifying customers is unstable or fluid.
You might be using a form of identifier, like a login or web cookie, that isn't unique (to a person or household) and causes people to have multiple identifiers. If the unit of identity changes over time, people might see multiple treatments. For example, a person might be randomly assigned to the treatment condition on desktop, but later could be randomly assigned to the control group on his or her mobile device. Thus, in the absence of cross-device matching, the same person can appear in multiple experimental conditions.

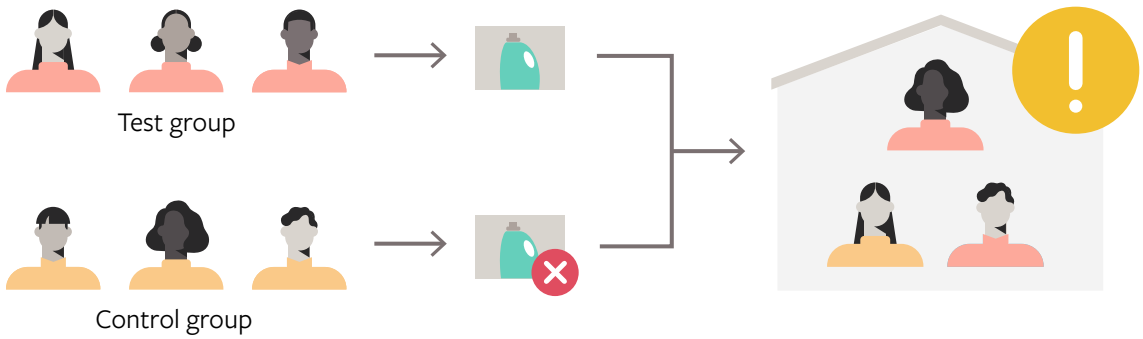
People in your control group talk to people outside the group.
Respondents may see your ad and tell their friends about the product or offering. This is especially common for entertainment or event-based outcomes. This word-of-mouth marketing means people who were not supposed to see the ad effectively still learn from it in a way that isn't measured in the test.

Someone else is making purchases for a test subject.
While you might split your treatments at one level, such as by person, some products are bought at the household level (for example CPG or insurance products). People in the household might be in multiple treatment groups and therefore may see multiple treatments.

Sometimes you get surprise variables.
There might be a fair split, but something could happen only to one of your groups, out of your control. An example of this could be when a natural disaster hits the treatment region in a geo-test.

What could this look like?

A CPG company that sells a variety of household goods is running a test to observe the effectiveness of online advertising for a new brand of laundry detergent. The treatment and control groups are split at the individual level. As the campaign runs, the marketing team quickly realizes that detergent is actually bought at the household level, and it is possible that individuals across the control and treatment groups reside in the same households.



What could you do?

When evaluating the experiment, it's important to understand how this can affect results and the scale of the effect. Generally, crossover between treatments dampens the observed effect, making the treatments look less incremental or showing a smaller difference between treatments.

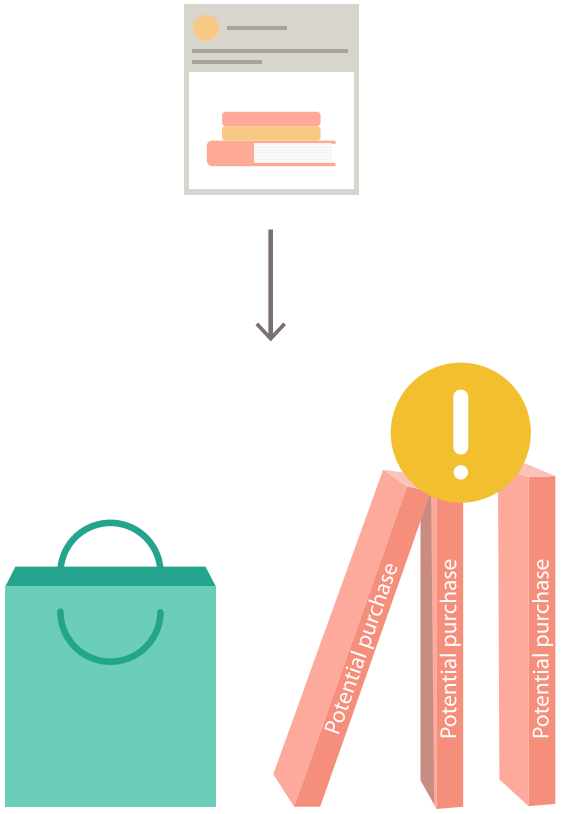
If you still see strong incremental results or a difference between groups, you can likely still use the results. They will just serve as a lower bound for the true impact of the treatment. If you think the number of people affected by the instability is too large to observe an effect, consider alternative testing options.

PROBLEM 5

Some Tests Will Have Effects Beyond an Initial User Interaction

When analyzing an experiment, one common problem is not taking into consideration the effects of a person's action after being exposed to an ad and what impact that has on other potential purchases. These are referred to as second-order effects.

It's crucial to evaluate these second-order effects beyond the direct user action you're trying to measure. For example, an action taken as the result of being part of a test can affect a customers' lifetime value, or what they do and don't buy in other related product categories. These scenarios can alter how you evaluate the success of a campaign.



How can this happen?

There are multiple scenarios where second-order effects can occur:

Tests can have a ripple effect over the lifetime value of a customer.
An experiment can affect the mix and expected value of your customers. Volume of future purchases, quality of customers or returns can all affect your perception of which treatment truly drives more value for your company.

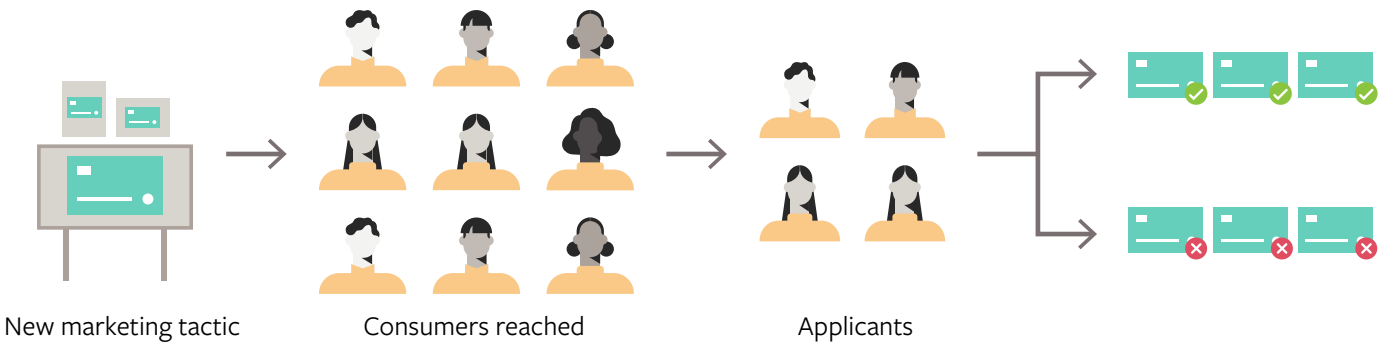
Tests can also affect other product categories.
Treatments can affect how consumers purchase other products in your portfolio. If you don't look at your company's portfolio of products, you might over- or undervalue a treatment.

Tests can lead a person to a particular purchase channel.
Some treatments can drive people to purchase from a specific channel, such as in-store instead of online. While this looks successful from one channel, holistically you may not be driving as much value.

Tests can help you convert customers you eventually would have landed—just sooner.
Your treatment could cause people who would have converted anyway to convert sooner. This is especially the case with promotions, vouchers and coupons. This can lead you to think that a campaign is more effective than it is in reality.

What could this look like?

A financial services company is testing a new marketing tactic to promote a credit card. Throughout testing, the marketing team gets nervous that the number of approved credit cards is looking very low, leading them to believe that the campaign is not effective. However, they quickly realize they failed to take into account that a certain number of consumers will not get approved for this card, even if they responded to the marketing message.



What could you do?

For tests like this, you should think through whether success is defined as an application or an approval, and build those KPIs into your test design and data assessment. The best way to handle second-order effects is to understand the assumptions you're making and directly measure downstream impacts to the extent possible while ensuring that you address any concerns around power. The truth is, even if a test reveals a tactic or campaign as a success, there are still going to be some things you can't measure. At most advanced measurement companies, every experiment is conducted with a suite of 100 or more metrics to check impact on other KPIs.

As a first step, understand your assumptions. Do you assume lifetime value is the same? Are you assuming that people purchase in the same channel? Have you thought of your top customer journeys so you can identify the assumptions you need to make? This helps you understand where your blind spots are or how to correct for them.

To the extent possible, you should measure these second-order effects during a test. For example, if you're trying to drive online purchases but can also look at offline sales, then measure both (although statistical power is likely lower for the second-order effects). If your transactions can have different values, you should evaluate order size or lifetime value in addition to number of transactions. If you think your treatment may cause people to convert sooner than they normally would, make sure you use a sufficiently long measurement period and look for metric changes to stabilize.

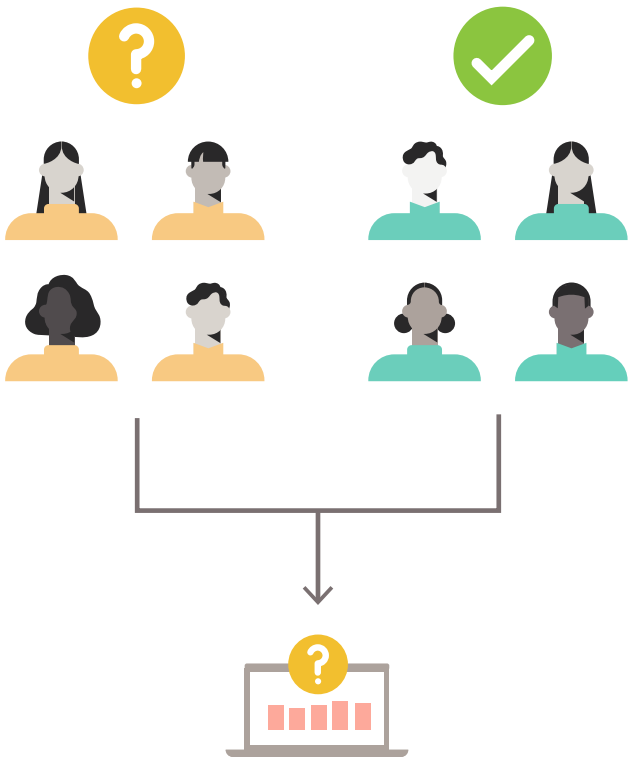
PROBLEM 6

You Can't Track Everything You Want

Just as some tests can be plagued by instability, others simply have unavoidable unknowns.

People may not always adhere to the platforms on which you are running the tests, and people don't always provide data when you want or need it.

For example, when running an experiment, you're reliant on knowing which treatment group a person is in when conducting the analysis. This can be difficult in places when outcomes occur in an anonymous fashion or on a different platform than the one on which the experiment is run.



How can this happen?

It's not uncommon for tests to be impacted by variables that are simply difficult to track:

People use cash, or make other untraceable purchases.
Some outcomes aren't able to be tied back to a person or other form of identity. One example is cash transactions in retail establishments that lack loyalty programs.

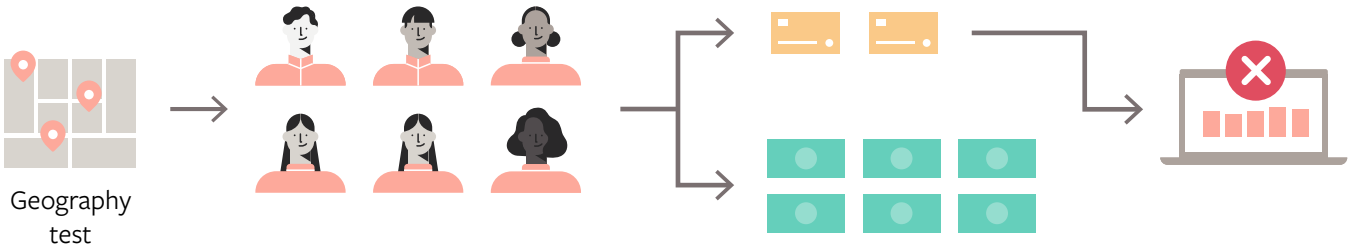
You just can't match up all of your user IDs.
When bringing data into a platform to measure, you may rely on a matching system that uses various identifiers. For many reasons, like the use of multiple emails or lack of common identifier for matching, you may not be able to match 100% of outcomes between platforms.

People's browser behavior makes matching difficult.
Matching outcomes to treatment groups online requires stable identifiers. User behavior, like device switching, or browser behavior, like deleting or blocking cookies, can make identifiers unstable.

You struggle to get enough people to respond to your brand study.
Some forms of outcomes, like brand impact, are best measured with polling. However, polling response, even among those prodded to respond, can be incomplete and therefore untraceable.

What could this look like?

A quick-service restaurant is running a geography-level test to understand the effect of media in the neighborhoods around it. However, as the test progresses, the marketer quickly realizes that the high percentage of cash transactions are very difficult to track as there is no data available to leverage.



What could you do?

First, it's important to understand how big of a problem this is for your business on two dimensions.

1. **What percentage of transactions are not traceable?**

2. **Is this percentage equivalent in test and control groups.**

If the percentage of untraceable transactions is low and this effect is even across treatment groups, you may be able to ignore the effects of traceability. This is especially true when making optimization decisions between treatments, as opposed to determining the true value or ROI of a strategy.

The most common way to overcome this is to perform an adjustment to correct for the lack of coverage. This allows you to see numbers as if there were no lack of traceability. This is essential for things like polling, but can also be applied to other outcomes. While you do have to make assumptions about the similarity between the traceable and untraceable population as well as match rates when factoring up, it can provide a truer sense of the scale of overall impact.

If the majority of transactions are untraceable, you may want to pursue experimental strategies that align with a higher-order traceable unit (the next best thing you can track if you can't track the metric you desire). For example if most of your transactions at your stores are cash-based, you can perform geo-tests that align to store-level traceability. Another example might be when a test struggles to collect data on individuals, a brand might opt to track regional impact, like across a certain zip code.

Conclusion

In testing, it is possible that things can go wrong, but that is expected as you continually iterate and learn. After all, it's why you run such tests—to learn and get better and smarter.

Because many of the problems that result from marketing tests are common and predictable, as we've seen in this guide, you'll know what to look for, and how best to respond when such problems arise. Plus, if approached correctly, even flawed tests can prove valuable. The most successful marketers encounter problems all the time, because they are constantly testing and learning—and getting better—along the way.

"I think that doing an experiment by itself does not ensure that you will get a perfect answer, because things do go wrong in the real world," said **Matthew Gerrie, Senior Director of Marketing Science and Communication, Booking.com.** "It's important to know what can go wrong and what you are willing to give up or ignore for the purposes of running your test, and what is so severe that you may need to start the test or restart. It's important to distinguish between those two—things that are bad that you can ignore and things that are so bad you have to stop."

By having a strong plan in place before you get started and anticipating some of the key areas that can go wrong, you will be setting your company up for success.

Alok Gupta, Director of Data Science, Lyft explained, "You shouldn't be afraid of things going wrong because ultimately, companies will reward responsible learning. Accelerating a company's learning is crucial."

The business benefit you will see from this approach will far outweigh any challenges you encounter along the way.