Why your AB Tests aren't Working

Core Challenges: A Cautionary Tale (Part I)

Authors: Donny Lofland and Christie Marsh, Ph.D

Introduction

Many digital media companies, marketing agencies, and software development studios run *AB Tests* to test the performance of new content and features, and optimize existing ones. Common examples include: "What shop button color drives more sales: red or green?" and "What store page description maximizes app installs: version A or version B?" In both cases, the reasonable answer seems to be "Let's run an AB Test to find out." If you work with consumer products and data analysis there is a good chance that you have been or will be asked to help with AB Tests.

On the surface, AB Testing is a great idea and seems relatively straightforward; however, we often develop tests from unsubstantiated assumptions, and as the tests become more nuanced, their conclusions become more dubious. Look deep enough, and it all unravels. Our goal is to share what we've learned from years of combined experience running AB tests in industry settings to help you identify the myriad challenges inherent to AB testing you may face that prevent meaningful and actionable results.

Key Terminology defined below article.

What is AB Testing (True vs. Best)?

Scientific Experimentation

Scientists use experimentation to generate new hypotheses and investigate the likelihoods that they are *true*. These experiments are conducted in highly controlled environments to ensure, as best as possible, that observations are attributable only to the variables being manipulated. This allows scientists to make inferences about causation, which requires three conditions (Kenny, 1979):

1. Covariance: when A changes, B changes too
2. Temporal precedence: B only changes after A changes first
3. No third variables: the only thing that could have caused B to change is A

AB Testing

In AB testing, we can typically only establish the first one: covariance. Strictly controlled environments are often required for the other two, but our users/consumers don't use our products in such environments and people behave differently when they know they are being observed. We want to understand how our products perform with unobserved people in real environments. AB testing thus *cannot easily establish what is true* so instead, we focus on *what is best*.

> "AB testing thus *cannot easily establish what is true* so instead we focus on *what is best*."

In a standard AB test, the user base is divided randomly into two distinct experiences, A and B. After some time has passed, we evaluate performance, typically in *retention, engagement, or monetization*, and close the test to the experience that resulted in the best performance. By way of example, let's return to our shop button. Let's say you ran an AB test for two weeks where 50% of your new users saw a ***green*** shop button and 50% of your new users saw a ***red*** shop button. At the end of your test, you evaluate click-through on the button (CTR) and determine that the red button resulted in more users clicking through to the shop (Figure 1).

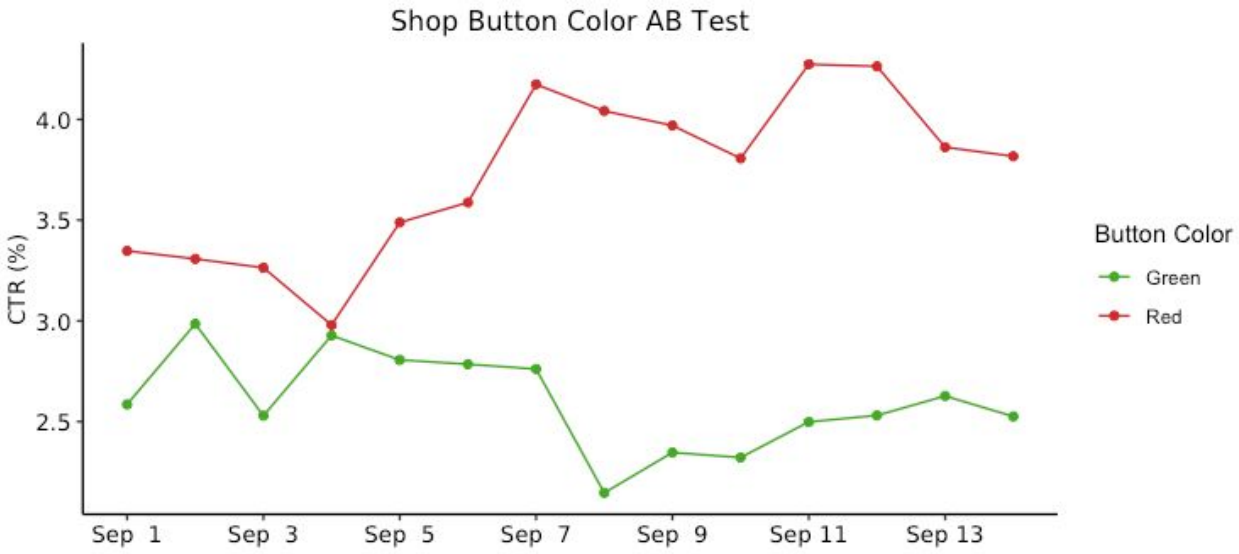


Figure 1: AB test results comparing the click-through rate (CTR) for a green shop button versus a red shop button over a two-week period. The red shop button emerged as the winner at the end of the test.

We don't necessarily care *why* performance was best for the red shop button, only that *it was*, but this is problematic. Our AB test did not help us understand what is *true*. We are unable to say *why* the red shop button resulted in more clicks and we typically don't wish to invest resources doing incremental testing to find out because the rate of return (ROI) from doing so is rarely obvious or easily justified; however, if we are unable to explain the *why* of our results then we are unable to know if the experience that was "best" was actually *best* or if it was the better of two bad options, spurious, *best* only for the current app/product version, or only *best* during the timeframe of our test.

This distinction is important. *True* implies some fundamental property that doesn't change, while *best* means we saw a possible transitory correlation or pattern between experiences and results. **As soon as we make any change to our product or our marketing strategy (bringing in a different audience), *best* can change right along with it, invalidating all past AB test results.** What does this mean for our shop button? We observed that red was best, but what if, for example, red was best only given the context of the color of an adjacent asset. Perhaps red was best because the adjacent asset was blue and red contrasts with blue more than green does making the shop button stand out more. If we ever change the color or nature of the assets adjacent to our button then the results of this AB test are moot. If we didn't understand what was driving the performance of the red button then we may lose revenue when we make app changes in the future that negate the positive impact of a red button on shop entrances.

"But our designers would know very well why red was optimal, that's why they chose to test it!", you exclaim. Or perhaps you did your research to make an informed decision on what to test in the first place; surely, being deliberate instead of taking shots in the dark negates the issues? These are both completely fair reactions; however, it is important to consider that human

intuition is frequently wrong, and products are complex and often have conflicting features that make interpretation far from simple and analogies to results found in research messy. Complex and difficult to predict interactions abound. Your intuition about the causes of observed effects is predicated on the context in which those intuitions were formed and they may differ considerably from the context of your app or product in unrealized ways. This is why Science is slow and incremental, a process that industry moves much too quickly for.

How we get it wrong!

“But I am not trying to be a Scientist or to do Science, I just want to know what color to make my shop button”, you think! And you would be right in doing so. We know we’re not trying to establish fundamental truths about human behavior and experience, so why does it matter that we’re not conducting controlled experiments in laboratory settings?

Changing or Unknown Users

When scientists in labs perform research, they typically have the advantage of a very controlled setting and known subjects. They can carefully manipulate single variables at a time to explore what happens when you make a change. In the consumer industry, however, our users are people we haven't chosen or targeted for participation in a specific experiment. They are not provided strict instructions for interacting with our products and thus do so in many unanticipated ways. They bring different backgrounds, biases, prior knowledge, resources, understanding, nuances, and expectations with them that can affect their behavioral decisions within your AB Test. To complicate the picture, users often don’t make rational or even consistent choices (Thaler, 2016). The field of Behavior Economic explores the patterns in which people often behave irrationally in surprisingly consistent ways.

Most consumer user bases are also constantly evolving, and there is no way to guarantee the future makeup will match the current makeup when you ran your test. Unlike carefully constructed experiments in academic research, we often have an incomplete picture of our users, and this cannot account for real factors affecting their decisions. We also aren’t investigating core human behaviors that evolved long ago, but human interactions in environments we didn’t explicitly evolve to exist in and with technology that has existed for a very short time in human history. Even in controlled experiments in labs, old, well-established theories are still often re-tested at the outset of a new experiment to demonstrate the foundations the new hypothesis rests on are still true.

This unknown and uncontrollable human component also extends to our environments. Besides variability across users, AB Tests are complicated by market shifts, your competition, different definitions for a “winner,” multiple stakeholders within your organization, etc.

Complex or Multiple Tests

Problems can arise when multiple tests are run in parallel on the same user base or more than one change is being evaluated in the same test. If we have multiple changes happening at the same time, we cannot attribute which led to measurable effects. We also run the risk that multiple changes can cancel each other out such that we don't see any overall performance differences when one change made things better and the other made things worse. In science, we try to isolate changes and only measure the effect of one variable at a time on our measured outcome. In business, we ideally want to run as few AB tests as possible and might be tempted to combine multiple things in a single AB test.

Sample Sizes

If you don't see a clear winner in your AB Test, you may be tempted to just run the test longer and collect more data until you do find a winner. Scientists understand that given a large enough dataset, "statistically significant" differences are virtually guaranteed; however, those differences are suspect and more likely to be meaningless or so small as to not effectively matter. There is a term for this and it's called "p-value fishing". When we run AB tests, we want to find **meaningful** results or it's not worth our trouble.

When designing tests, you need to determine upfront whether you will be able to actually detect differences and if not, rethink your test. You should use a sample size calculation to estimate how many users you will need in each group to detect some threshold difference with some confidence level. Then, run your test long enough to gather those users and stop. If you don't find a statistical winner, then call it a tie and move on to another question.

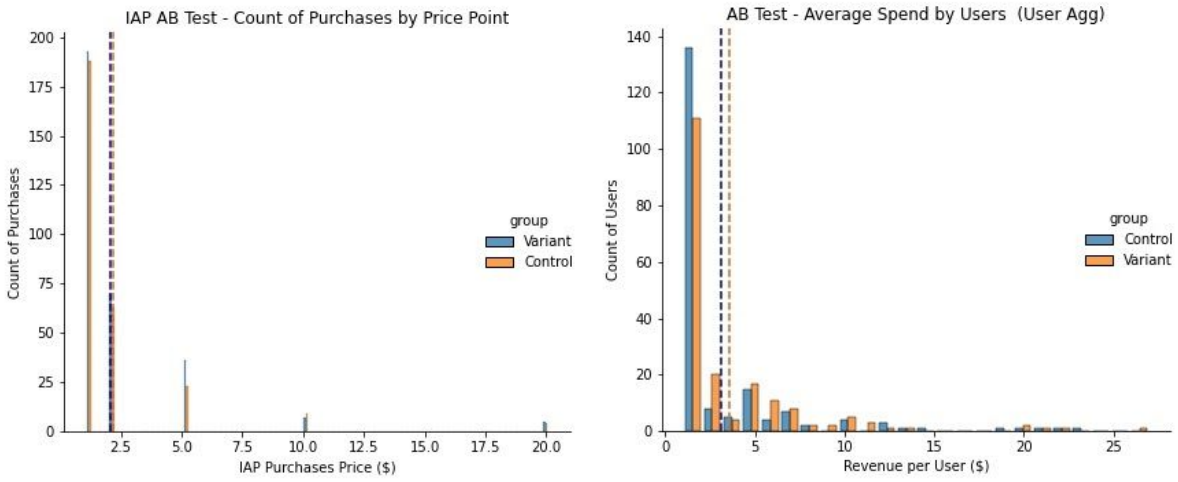
Hunting for a Needle in a Haystack

Sometimes we are hoping to detect changes in a behavior that doesn't happen often which can be especially challenging. For example, you might want to run a test and measure the difference in spending between two groups, but maybe you have a very low conversion rate. It's quite common with mobile apps to see conversion rates under 2%. Let work through an example.

> **Setup:** You set up an AB test with the goal to collect **20,000 users**. You end up with **9967 users** in your *Control* and **10033 users** in your *Variant*. You've made a change in your *Variant* and are hoping to improve the average in-app spend by users. Users can purchase items for $1, $2, $5, $10, or $20. From historic data, you only expect ~2% of users to make a purchases and most users make smaller dollar purchases.
>
> **Results:** Out of your *Control* group, you had 287 purchases by 191 users and a total revenue of $599 (mean: **$3.14**, standard deviation: 3.95). In your *Variant,* you ended up

with 311 purchases by 190 users for a total of $683 (mean: **$3.60**, standard deviation: 4.08). ***Did your Variant win?***

Let's deal with the first gotcha … What does the distribution of purchases and spenders within each group look like? In our experience, spending patterns are most commonly skewed, longtail curves with the majority of users spending very little and a few spenders spending quite a bit. These curves are rarely normal and often have a more exponential shape. Here is what our data might look like. In the first chart, we graph the count of each purchase by it's price and in the second. Since users can make multiple purchases, in the second chart, we tally up the total spend per user. Note that the **blue** dashed line represents *Control's* "Average Spend" and the **orange** dashed line represents the *Variant's* "Average Spend".



Notice how the blue and orange "average" lines are closer to the lower $ amounts (more towards the left of the graph) and there are fewer users in the higher $ amounts? Also, you get different averages when you aggregate spend to each user (accounting for multiple purchases by users)? Both these curves are skewed with more users in the $1~$5 range and a fewer greater than $5. When we are working with skewed data, means (averages) don't tell us where the bulk of the are located. In fact, when working with heavily skewed data, other KPIs like the *median* or the *geometric mean* are better summary statistics to use for comparison. A challenge with using a *mean* is that it is far more sensitive to those few high values. If we were to repeat this experiment over and over, we cannot rely on getting those same high spenders every time - as a result, your mean is less stable and less trustworthy.

That said, most business analysts love the mean because it's such a simple intuitive value. Take your total revenue divided by total purchases or users and viola, a simple number for comparison. However, any conclusions you draw from using a mean with skewed data are less likely to be real and you are more likely to pick an incorrect winner.

From this data, the means we arrived at are: *Control* **$3.14** and *Variant* **$3.60**. Looks great, Right? *Variant* wins? Not so fast, you cannot draw conclusions for means and very skewed data!.

That said, let's use our imaginations and pretend for a moment that your users did fit a nice normal distributions where it's ok to use the mean to compare the groups. We still have a problem. Let's see what happens ...

Starting with just conversions, a 1.92% conversion rate (191 spenders out of 9967 users) is not significantly different from 1.89% (190 spenders out of 10033 users). Calculating statistical significance (using any number of online conversion calculators, e.g. SurveyMonkey AB Test Calculator) we end up with a p-value of ~0.54. So we cannot say with any confidence that our change led to a change in purchasers. That said, *maybe the variant really did lead to fewer spenders … we cannot tell with such a small percentage of users!*

Next, let's consider the mean spend by users in the two groups. If we use a comparison of means calculator (widely available online, e.g. MedCalc Comparison of Means) and enter our means, standard deviations, and counts, we again find there is no statistical difference and a p-value of ~0.26. While the means looked different and we were hopeful the *Variant* won with $3.60 (over $3.14), that difference was more likely due to random chance. But, here's the kicker - *it might have actually been a winner but we just couldn't tell with such a small sample size!*

We've run into a classic problem with AB testing - the signal-to-noise ratio. Given the random variability, we cannot reasonably tell differences with such small changes between groups and small sample sizes. Yes, we can use a sample size calculator to figure out how many spenders we might need, but this number could be in the hundreds of thousands to millions depending on the difference we are trying to detect. Yikes!

If you cannot detect your differences with the desired KPI (e.g. Average Spend per User), you might still be able to pick a winner using other criteria acting as a proxy for potential spending (e.g. click-thru-rates or progress in your app). In your dataset, you have another ~9,810 players in each group who didn't spend. While you didn't see a statistical difference in IAP in your experiment, maybe CTR suggests the *Variant* is more likely to lead to purchases and thus is a good choice.

Incorrect or Expiring Conclusions

The unfortunate truth is that *for any AB Test you run*, no matter how well thought out or how significant the results, *your results and conclusions may not be true even if they appear best, and even if they are genuinely best, they have a shelf life*. In the best case, your "winner" is still the favorite for a long time, but in the worst case, not only can your "loser" become (or has always been) the "best" variant, but it can do so soon after you close your test.

> "The unfortunate truth is that *for any AB Test you run*, no matter how well thought out or how significant the results, *your results and conclusions may not be true even if they appear best, and even if they are genuinely best, they have a shelf life*."

When running an AB Test, your results are only good for the current conditions, including user base and version of your app or content. The moment anything changes, you can no longer be confident that previous winners are still your best option unless you rerun those AB tests.

What does this mean for us?

- Results from previous AB tests are possibly invalidated when you release new versions or make changes.
- Market shifts affecting the distribution of your user base, user expectations, and even features offered by your competition can all affect how users interpret and respond with your product. The variant previously determined to be the winner, if tested again, might not win next time.
- If you rely on User Acquisition (UA) or marketing campaigns to drive a sizable portion of your user base, you may start getting different quality users or users with different behaviors independent of your content or service. While random assignment to groups during a test should distribute the users and mitigate effects, different mixes of users can affect the variability or spread of measured KPIs, making it harder to identify a winner.

What we get wrong:

> ***We treat AB Tests as simple, stand-alone experiments guaranteed to have a winner without establishing sound, theoretical reasons a difference between variants should exist, taking into account the complexity inherent with running AB tests, or the rigor of ensuring differences we see are even real and not artifacts.***

Conclusion

As we've seen, there are many subtle challenges to designing, running, and analyzing AB Tests. This article isn't to say AB Testing cannot or should not be performed, but rather, you should be aware of challenges when designing an AB Testing Framework. If you are embarking on AB testing or have an existing AB test program, here are some tips that might help you find more success.

Tips for Success

1. Plan, plan, plan - the more thought you put into a test before you run it, the less work you'll face during analysis, and the better the chance you'll find meaningful results.
2. Treat AB testing as a short-term optimization problem - if a question is critical or shows meaningful results, plan to circle back later and see if it's still a winner.
3. Simple dashboards and simple metrics in most cases will not tease out meaningful results. AB test analysis requires more sophisticated statistical techniques.
4. Be willing to accept that sometimes there is NO winner. In those cases, use your instinct to choose what you feel makes the most sense in the context of other changes you might want to make or other tests you might want to run. You are the domain expert and your intuition probably performs better than chance. If you are not seeing "winners", then ask whether you should be AB testing that aspect of your product or whether you should focus attention on other areas where you might have a better chance to move the needle.
5. If you aren't doing segment analysis yet, consider exploring that direction. Find the subgroups that are most valuable or that have the most potential for improvement. For these groups, think about AB tests that might improve their experience and pay less attention to parts of your user base where you don't see value or potential.

In this article, we fleshed out the core challenges with AB testing at a high-level. In the next part, Seven (Not So) Simple Steps for Experimental Design, we will take a deep dive into the issues that you may face at each step of the AB test process.

Terminology

Here is a list of terms we'll use during our discussion. These are commonly seen in marketing, especially when talking about customers or users.

- **App** - "Apps," short for applications, but could be websites, web applications, mobile apps, desktop apps, video games, or any interactive interface where users interact and make choices. When we use the term app, it could refer to any product or service offer.
- **Churn** - When a user leaves your app, never to return. Churn could also mean a customer that won't return to a retail establishment.
- **Cohort** - Typically, a group of users that all started on the same calendar day, for example, installing a mobile application or signing up for a service.
- **Engagement** - Any measure of how much someone is interacting with your service, app, or content. For example, the *number of views*, click-thru-rate (CTR), *time spent reading*, *time in-app*, or *progress through fixed user experiences*.
- **KPI** - This stands for "*Key Performance Indicator*" and is something we measure and use to quantify users or groups in some way. Common KPIs can be related to the user

value (e.g., *lifetime value (LTV)*, *average revenue per user*), **engagement** (e.g., *in-app time*, *sessions per day*), **retention** (e.g., *days since install*), or any **app-specific measures** (e.g., *the level reached*, *screens visited*, *links clicked*).

- **Retention** - How long a user remains a potential customer or sticks around in your application or service.
- **Variability** - This is a measure of how much random noise is in your data. With any experiment, we expect there to be random factors outside our control. If everyone behaved in the exact same way, we'd expect little to no variability. However when running tests with real people, you will see a mix of player types making different choices.

- **Control** - The specific variant group that has seen no changes. This group serves as a baseline for comparing other groups. If you are doing something for the first time and don't have a baseline or control group, you can arbitrarily assign one group to be your control.
- **Variant -** Each distinct group of users in an experiment. If an AB Test consists of 4 groups, each group is called a *variant*. Typically, each group has a slightly different experience and we are trying to measure outcomes between groups.

References

Kenny, D. A. (1979). *Correlation and causality*. Wiley-Interscience, New York.

Saffron, Paul. 2007. "Six Rules for Effective Forecasting" Harvard Business Review, July–August 2007 Issue.

Thaler, Richard H. 2016. "Behavioral Economics: Past, Present, and Future." American Economic Review, 106 (7): 1577-1600.