A/B Testing

SAILTHRU
Best Practices

Always Be Testing

The "always be testing" bug runs deep here at Sailthru, and we encourage our clients to follow suit. That said, all too often A/B tests are run in haste at the aim of driving quick wins, which can often have negative long-term effects. Be sure to think about these tips, tricks and best practices before you deploy your next round of testing.

Statistical Significance

Statistical significance is at the core of a successful test plan, but all too often it is forgotten. *You cannot have confidence (pun intended!) in your optimization strategies if you have not achieved statistical significance.* The fastest way to statistical significance is either sample size or a large delta in the metric you are testing (which is not alway so easy to achieve!). If it's been a while since Statistics 101 and you need some help calculating significance or estimating the sample size you'll need for an effective test, we've put together a one-click cheat sheet to help you out; contact your account manager for details.

Controls

The most valuable test results are the ones that can showcase *incremental* gains. For this reason, it is important to leverage holdout/control groups. Thinking about a retail example: you want to test a win-back campaign to disengaged customers and you want to know how many more conversions you'll get when you offer an incentive. To do this, you A/B test a win-back with an incentive vs. one without. WRONG! We are missing one important test cell here which is the holdout group. Believe it or not, there is always some element of chance that these disengaged customers will come back on their own. For this reason, we must have one test cell that receives no email at all and then the two previously discussed. The holdout group can be smaller if you are obsessing over revenue so long as you're sure the sample sizes will yield significance.

Pick Your Battle

...and just *one* battle. Sailthru's core platform is designed to support A/B testing. While multivariate testing is possible, you'll need to be well-versed in how to conduct a multivariate test. Why does this matter? In an A/B (or A/B/C/D/E...) test, we are testing one variable. Let's say you're designing a new email layout and wanted to include Sailthru Horizon-powered content in that new format (and you hadn't used these recommendations previously); you would first need to test the new template vs. the old (or alternatively, the old template without Horizon vs. the old template *with* Horizon) and then proceed to a second test. If everything is done in one fell swoop, you will be unable to discern the lift yielded by the new design from the lift from the recommended content. Another option would be to run an A/B/C/D test: A) old template, no Horizon; B) old template, Horizon; C) new template, no Horizon; D) new template, Horizon.

Campaign Testing vs. Cohort Testing

Whereas your standard split test might focus on driving incremental revenue through a simple tweak (e.g. testing subject lines or testing free shipping vs. $20 off vs. 10% off in a welcome email to see which offer yields the stronger revenue per send), cohort studies are designed to look at the impact of different treatments *over time.*

Let's stick with something along the lines of the aforementioned example: seeking to optimize its welcome stream in the name of incremental conversions, a retailer tests a 10% discount offer to new subscribers at various points in the first 14 days. Not surprisingly, the promotional offer results in a lift in gross conversions (Marketing 101: promotion moves product!). After digging one level deeper into the numbers, the retailer notes that the offer also prompted an increase in average order value (AOV), likely due to classic stockpiling effects. Sounds like we've found a winner in the promotion takers, right?

No, we have not. Instead, we find ourselves with several follow-on questions around the downstream impact of this promotion – namely, did the early discount offer train the customer to buy on promotion and erode downstream lifetime value?

Dissecting the Surface Metrics

Continuing with that example, perhaps the marketer is particularly in tune with the numbers and actually analyzes some downstream numbers several months after the initial welcome stream. After the analysis they note that AOV remains higher for those who initially converted with a discount offer. With this new data point, it seems as though we have finally identified a solvent winner in the discount cell, have we not?

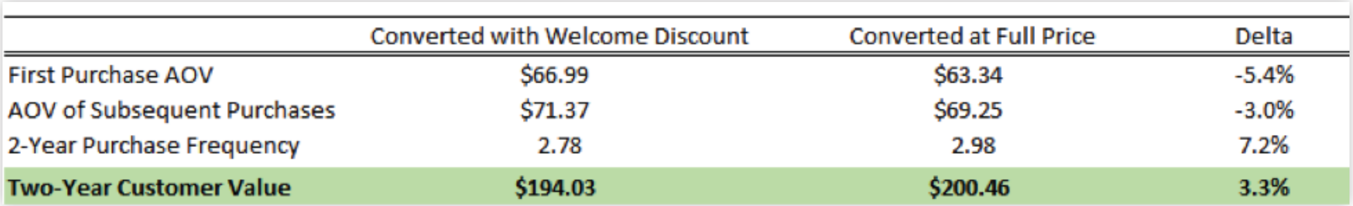
In fact, we unfortunately still have not. Consider the chart below, which now also takes into account the purchase frequency of the two different cohorts over two years. Despite the promotion takers yielding transactions between 3% and 5.4% more valuable than those customers paying full price, they purchased at a rate of 7.2% less than those full-price shoppers (again, could be a function of stockpiling), meaning they netted out to be 3.3% less valuable than those with the seemingly more expensive carts.

Even with this revelation, though, it's important to revisit the QQQ: the quantity/quality quandary. Are there enough non-discount shoppers at that higher two-year value to keep gross revenue compelling? If the number of buyers falls considerably when that promotional incentive is removed, the marketer will need to think twice.

Short-Term Experiments vs. Long-Term Learning

Recently, many Sailthru clients have leveraged cohort-based tests to assess the impact of email frequency on subscriber opt-out rates. More specifically, they use customer-level variables to dictate two groups: one that receives emails daily for the first 60 days and another that receives only three messages per week. From there, they can compare/contrast 60-day opt-out rates and other engagement metrics to understand the potential impact of tweaking frequency.

Campaign-centric testing – tweaking elements such as subject lines and calls to action can certainly be valuable for driving incremental ROI, but it's mission-critical to double-check how you think about testing. Rather than running separate A/B tests on your welcome email, your day 2 email, your day 7 email, etc., should you be developing welcome series A vs. B and ensuring that one cohort receives ALL A cells and the other ALL B's (read: cutting the test groups at the customer level vs. the campaign level)? Probably so.



	Converted with Welcome Discount	Converted at Full Price	Delta
First Purchase AOV	$66.99	$63.34	-5.4%
AOV of Subsequent Purchases	$71.37	$69.25	-3.0%
2-Year Purchase Frequency	2.78	2.98	7.2%
Two-Year Customer Value	**$194.03**	**$200.46**	**3.3%**