The Essential Guide to A/B Testing for Digital Advertisers

adacus

Table of Contents

INTRODUCTION: If It's Broke, Fix It 3

CHAPTER 1: Get More From Less 4
Reduce the number of required impressions by using modern statistics

CHAPTER 2: Take a Closer Look 8
Drill down into segments to discover which messages work for whom

CHAPTER 3: Avoid Common Pitfalls 9
Design tests that will translate into ongoing sales lift

- Impression-Level Tests
- Testing within a Rotation
- Users Split by Line Item
- Testing Plan Siloed from Media Plan
- Creative Test Obscured by Media Attribution
- Missing Conversions

CONCLUSION: If Creative Matters, then Creative Testing Matters 12

adacus

Introduction: If It's Broke, Fix It

As creative is growing in importance, so too is creative A/B testing. A/B testing is the digital feedback loop that powers agile marketing and continuous creative optimization. Despite its growing importance, creative A/B testing is largely broken, as most tests simply don't translate into ongoing sales lift. This is an enormous untapped opportunity for digital advertisers who understand proper testing.

At the same time, digital marketing poses unique challenges to the use of A/B testing. Failure to understand these challenges costs marketers millions of media dollars on A/B tests that simply don't lift performance.

- Most creative is completely untested because of faulty assumptions about sample size requirements for statistical significance.
- When advertisers *do* A/B test their creative, poor test design (using clicks or testing creatives within a creative rotation) prevents results from translating into ongoing sales lift.
- Advertisers often prematurely dismiss apparently insignificant overall results, instead of drilling down into audience segments to discover what messages work for which segments.

Fixing A/B testing in digital marketing has never been more urgent. Creative testing mattered less when impressions were competing with numerous placements on low-value inventory or were unviewable, because creative itself mattered less then. But with the higher quality that brands are increasingly demanding, and the leaner ad experiences being rolled out by publishers, advertisers are finally getting the viewability and attention to their ads they have longed for, and must not squander customer attention with untested creative.

As creative grows in importance, traditional copy testing just won't do. Pre-testing with a panel is slow, costly and not reflective of the constantly changing media contexts in which ads are experienced. The pace and volume of creative production demands continuous delivery of creatives, leveraging mid-flight feedback to optimize creative and creative plans.

Here's what brands need to do to fix A/B testing in digital marketing.

1. Reduce the number of required impressions by using modern statistics
2. Drill down into segments to discover what messages works for whom
3. Design tests that will translate into ongoing sales lift

Chapter 1: Get More from Less

Reduce the number of required impressions by using modern statistics

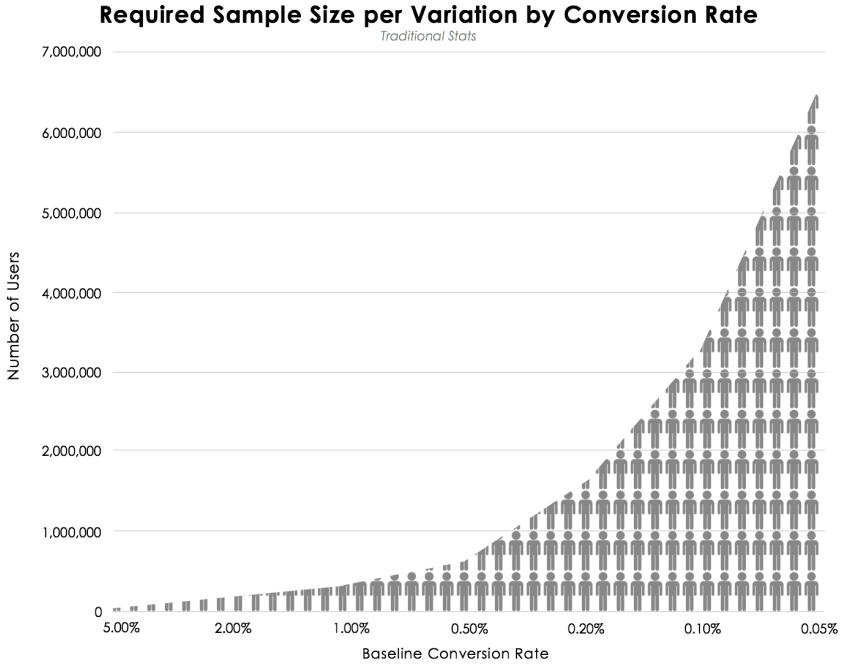
Many digital marketers hoping to run an A/B test are stopped in their tracks once they learn how long it will take to get results. The result: most creatives go completely untested.

A/B testing became ubiquitous in marketing with the explosion of e-commerce. Brands that sell online A/B test their landing pages, home pages, product detail pages and shopping carts constantly. No design element is too big or too small to A/B test, from the hero image and site hierarchy to button colors and fonts.

But there's one big difference between web sites and digital advertisements: the conversion rate. Web sites are generally lower in the sales funnel than digital advertisements, which means that their conversion rates are much higher.

In addition, web site A/B testing often optimizes against click-through rates, which have an even larger "conversion" rate than orders. In digital advertising, on the other hand, A/B testing against clicks is not advised, given the high rates of click fraud and accidental clicks. A/B testing of digital ads should generally be done against view-through conversions (orders or site actions) in order to ensure test results translate into ongoing sales lift.

Unfortunately, as conversion rates get smaller, the required sample size only gets larger. And the increase is not linear, it's exponential, as can be seen in the chart below.



adacus

But where do these sample size requirements come from? First generation A/B testing platforms reused the statistics that has traditionally been used for offline hypothesis testing, such as clinical trials. Traditional hypothesis testing statistics is based entirely on calculation of a p-value as the sole measure of significance, as explained below.

A new generation of A/B testing statistics is being developed by online optimization platforms such as VWO[1] (website A/B testing), swrve[2] (mobile app A/B testing) and Adacus (digital advertising A/B testing). This new generation of A/B testing statistics is based on Bayesian statistics, which is increasingly leveraged for much of today's data science and machine learning.[3]

What is "statistical significance"?

Marketers perpetually ask if test results are "statistically significant," but what does that term even mean? Probably not what you think.

Traditional statistics relies on p-values as a measure of the "statistical significance" of test results. Most marketers are surprised to learn that p-values do not, in fact, measure the probability that one creative or treatment will outperform another. What p-values measure is far more abstract and removed from the decisions that marketers make based on A/B tests.

In every A/B test one variation will perform at least slightly better than the other. P-values measure the probability that a test result (say, creative variation B outperforms creative variation A by 10%) would have occurred if in fact there were no difference between the two creatives at all. That "95% confidence level" threshold you've probably heard bantered about simply means that there is a 5% chance that, were the two variations identical, you would have observed as large a difference in performance between them as you did. The p-value is an important measure in other fields of study to account for what is known as in traditional statistics as Type I error.[4] In our experience, we have yet to hear a digital marketer ask us for this specific probability. And why would they.

Bayesian statistics makes use of two metrics that are critical to making decisions based upon A/B tests: Chance to Beat and Potential Loss.

- **Chance to Beat Control:** Most marketers assume that p-values reflect the chance that one creative variation will beat another. As described above, they only measure one source of sampling error. Chance to Beat Control, however, is based upon direct analysis of the probability distribution for each creative variation, and thus answers the question being asked by marketers directly.
- **Potential Loss:** You've selected a winning creative variation, and you know based on the Chance to Beat that there is a chance it's a false positive and the losing variation could beat it. But do you know by how much? You learn this from the Potential Loss of selecting a variation as the winner. This tells you how much is at risk if the losing variation turns out to be the winner. In other words, if you were to rerun the test an

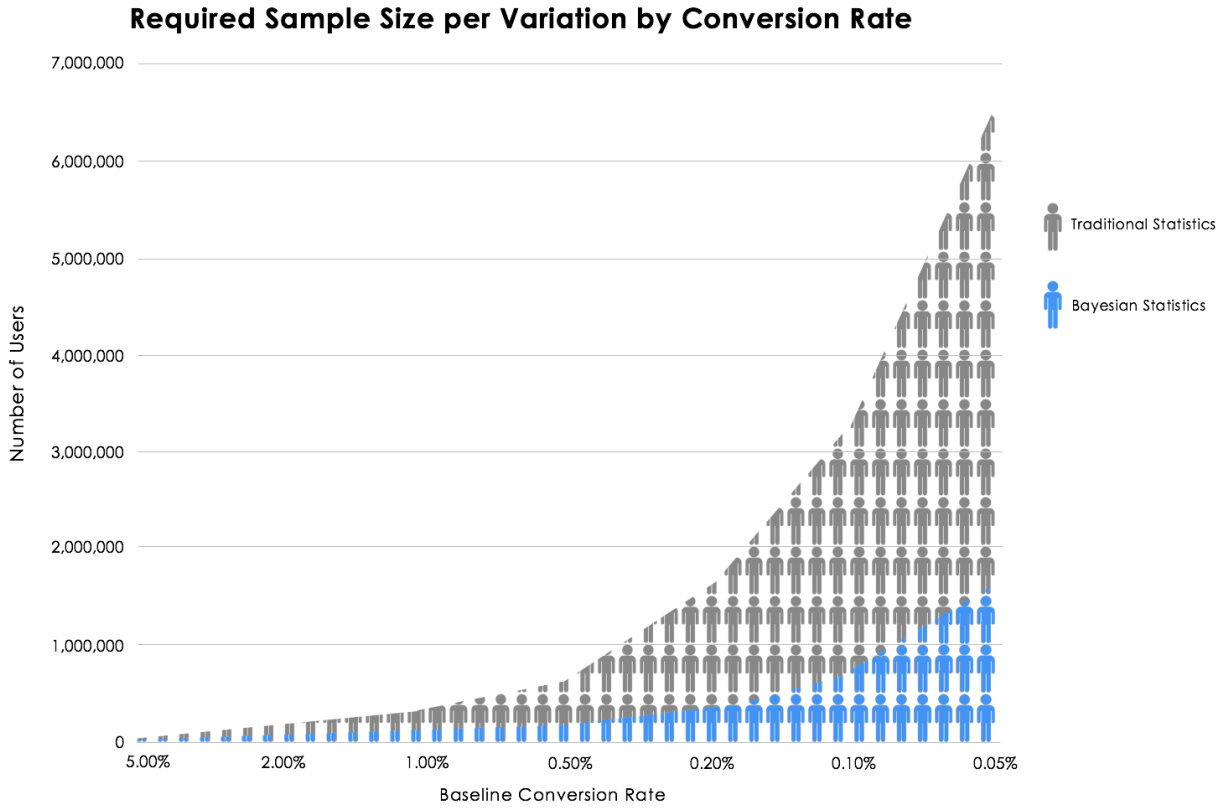
infinite number of times, and note every time the selected variation lost, this metric tells us the average loss in all those tests.

So, while p-values draw an arbitrary line in the sand, a line based on an abstract and non-intuitive measure of significance, Bayesian A/B statistics provide substantive, actionable measures of significance.

"Bayesian hypothesis testing is…on the cutting edge of things right now."[5]

-John Kucera, data scientist, Adobe

And not only do these metrics answer the actual questions that digital advertisers are asking, they are knowable with far smaller sample sizes than required by P-values.



As the chart demonstrates, the difference in sample size requirements between traditional and Bayesian A/B test statistics gets larger as conversion rates get smaller. For creative to be tested and optimized at the fast pace of digital advertising, Bayesian statistics are a critical tool in the digital marketers' toolkit.

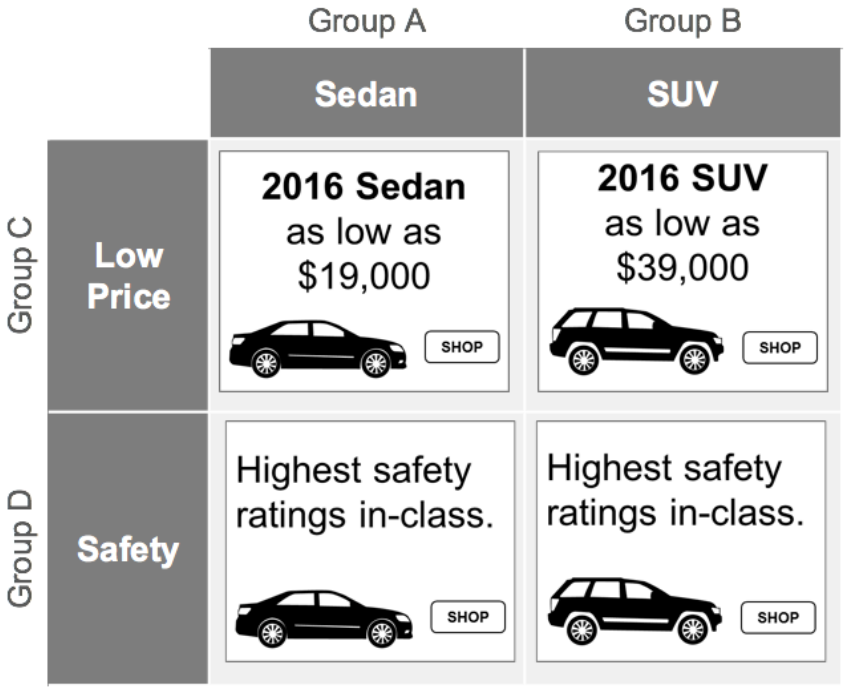
Get Results Even Faster with Full Factorial Test Design

In combination with Bayesian statistics, full factorial tests can also be used to reduce the required sample size of a digital advertising A/B test, thus speeding up time-to-insights and enabling mid-flight creative optimization.

Full factorial tests consist of two or more "factors" (elements to test) each of which has multiple options. This allows marketers to run multiple A/B tests simultaneously without increasing the required sample size.

While generally used for multivariate testing – to identify the effect of combinations of variables – the results of full factorial tests can also be analyzed for each separate factor in the test. This enables you to run multiple A/B tests concurrently, reusing the same sample of impressions.

In the below example, full factorial test design allows an automobile advertiser to test vehicle models (Group A v B) and messaging (Group C v D) simultaneously.



Chapter 2: Take A Closer Look

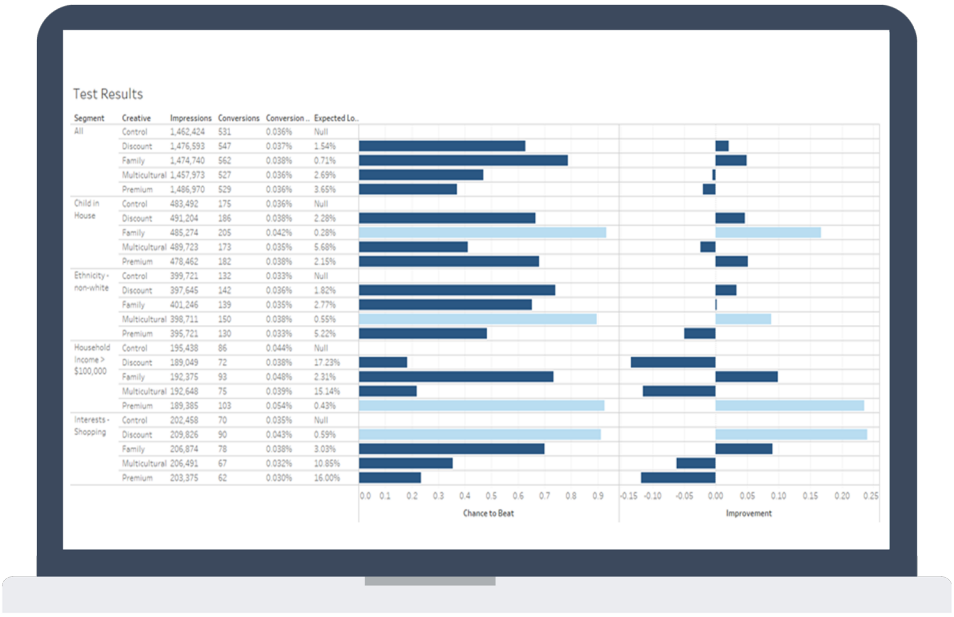
Drill down into segments to discover which messages work for whom

Many A/B tests that appear inconclusive are hiding their insights just beneath the surface. A creative variation that doesn't outperform overall may in fact outperform within a gender, ethnicity, location, household type, on a publisher's site, or within any other segment.

Obviously not all customers buy a product or service for the same reason. For brands to grow, they must internally differentiate themselves for new market segments. For digital marketing to drive growth, brands must communicate their unique value propositions differently to these different segments.

That's why it's critical to drill down into A/B test results to see which creative variation is the "winner" within various audience segments.

Of course, drilling down into A/B test results for a audience segment means looking at a smaller sample size, which makes the use of Bayesian statistics even more critical.



adacus

Chapter 3: Avoid Common Pitfalls

Design tests that will translate into ongoing sales lift

Even with the right statistics, however, most A/B tests are designed in a way that all but ensures the learnings will not translate into an ongoing lift in sales.

Ultimately, to know with confidence that one creative variation will outperform another, tests must be deployed at the user level, and not the impression level. When tests are run at the impression level, results are far more likely to indicate no difference between creatives. That's because impressions don't engage with brands, people do.

Impressions don't engage with brands, people do.

The large number of inconclusive A/B tests resulting from testing impressions rather than users unfortunately discourages many advertisers from continued testing and optimization. User-focused digital marketers are smarter, however, and demand creative A/B testing that is administered at the user level.

Impression-Level A/B Tests

Impression-level A/B tests occur when a marketer administers an A/B test by randomly assigning each impression to the A or B group. This creates the following problems:

- Users see both creative variations.
- When a user makes an order or other measurable brand engagement, one must use an arbitrary rule to determine which creative variation viewed by the user is credited with the conversion.

Such tests are often inconclusive even when the creative variations do perform differently.

To measure the actual difference in performance between creative variations, it is critical to randomly assign users, not impressions, to different groups and to serve the same creative variation to a user throughout the duration of the test.

A/B Testing within a Rotation

Most commodity ad servers offer a feature within creative rotations that optimizes the percentage weight of each creative within the rotation based on conversions as measured by clicks or by on-site events. Digital advertisers are strongly advised to avoid this feature and instead conduct actual A/B tests.

A/B testing via rotation optimization is a form of impression-level testing. It requires ignoring the impact of creatives that are not the last creative viewed or clicked, which negates decades of research on how advertising shapes buying patterns.

Ironically, optimizing weighting of creatives within a rotation is premised on the assumption that rotations are bad, because the impact of creatives that are not the last viewed or clicked within a rotation is presumed to be zero.

This is not to say that creative rotations aren't often better than serving a single creative to a user. In fact, a common A/B test is to test the impact of serving the same creative multiple times versus serving a rotation of creatives to a user.

Users Split by Line Item

Tests are all-too-often conducted by comparing the performance of two campaigns or line items being run simultaneously. The problem with this approach is that it does not hold the audience constant, which is a fundamental requirement of A/B testing. Different line items, even with identical settings, will inevitably access slightly different inventory over the course of the test. A proper A/B test in digital advertising must be run on a single line item with the ad server assigning users into groups at each impression so that audience used for the test truly is identical across the test groups.

Testing Plan Siloed from Media Plan

Finally, it is critical that creative A/B testing is coordinated with your media agency or in-house programmatic team. Programmatic media buying that is not coordinated with creative optimization can undermine creative tests in the following ways:

1. *Buying low CPM inventory that is less viewable limits the impact of any creative*

   As discussed in the introduction, creative is growing in importance for multiple reasons, one of which is that digital advertisers are increasingly getting the viewability and attention to their ads that they have been missing. Your creative testing will only be as informative as your ads are viewable. Ad placements that compete with 5-10 other placements on the page may be low in CPM, but the attention garnered makes creative less effective, thus making creative testing less effective. If you're impressions are 40% unviewable, 40% competing with 5-10 other placements, and only 20% both viewable and prominent, the impact of creative will simply be minimal.

2. *Targeting the same users in the A/B test with other programmatic campaigns*

   Sometimes when an agency sets up a creative A/B test, they generate a separate placement in the ad server for the test and traffic it to a line item or package with their trading desk. Such tests are less likely to measure the actual differences in performance between creative variations, because users in the test are being served creatives from other programmatic campaigns at the same time. When providing multiple placements or ad tags to a trading desk, ensure that the trading desk isolates users in a test from other programmatic campaigns.

Creative Test Obscured by Media Attribution

Ad servers, other than Adacus, apply attribution models by default to all conversion reporting. In other words, when a user converts after having seen ads from a test as well as ads from other placements or channels, the ad server will likely not attribute that conversion to the test group to which the user was assigned.

This makes A/B testing all but impossible, for two reasons:

1. Attribution across channels removes most conversions from the results of an A/B test, thus increasing the amount of testing time required to achieve significance from weeks to months.

2. Multi-channel attribution introduces noise into the A/B test results. Evaluation of an A/B test does not require multi-touch attribution as users are only presented with the A ad or the B ad throughout the duration of the test.

The solution is to report creative test performance separately from media performance across channels. In the creative test report, include all users that converted after seeing an ad in the test regardless of whether or not they were exposed to media from outside of the test as well.

Missing Conversions

When all of a digital marketer's conversions are found in online activity – site engagement, eCommerce orders – then conversion tracking is as simple as placing a pixel on their web site, but most marketers don't have it that easy.

Marketers whose businesses convert customers in call centers or in brick-and-mortar retail locations often unnecessarily limit their A/B tests to only the small subset of orders that are placed online. This significantly increases the required sample size of impressions, making A/B testing all but impossible for most digital advertisers.

The solution is offline conversion tracking. If your orders are placed by phone, you can leverage existing call intelligence technology to tie a call to an online device. If your orders are placed in physical stores, you can leverage sales measurement vendors to tie a purchase to an online device. To maximize your A/B Testing dollars, make sure your creative optimization vendor integrates with these offline tracking companies.

Conclusion

If Creative Matters, then Creative Testing Matters

While creative A/B testing is broken right now, fixing it represents an enormous untapped opportunity for digital marketers who understand it. A/B testing is the digital feedback loop that powers agile marketing and continuous creative optimization.

By leveraging the power of programmatic to gain insights, A/B testing enables digital marketers to deploy creative fast and make mid-flight adjustments and optimizations, all while learning more about their market and their response to different messages.

References

[1] See the paper, "A/B Testing at VWO", by Chris Stucchio, CTO of VWO. https://cdn2.hubspot.net/hubfs/310840/VWO_SmartStats_technical_whitepaper.pdf

[2] See "Why Use a Bayesian Approach to A/B Testing?" https://docs.swrve.com/faqs/resource-a-b-testing/bayesian-approach-to-ab-testing

[3] For an accessible introduction to Bayesian statistics, see "The Signal and the Noise: Why So Many Predictions Fail, but Some Don't"; Silver, Nate, Chapter 8

[4] For more detail on the limitations of p-values, see "What to believe: Bayesian methods for data analysis", John Kruschke, http://www.indiana.edu/~kruschke/articles/Kruschke2010TiCS.pdf

[5] "We currently use statistical test techniques that are based on what I would call classical A/B testing. It was developed in order to be able to measure and make decisions quickly in relatively low computing power environments. For example, most medical tests are still done using these techniques—the same techniques that we use in Adobe Target now—it's called "a frequentist approach" and it essentially assumes that you can repeat an experiment over and over again, and get some idea of what you might expect to see in terms of how much the results would differ. *We're looking at moving towards interpretation in terms of what's called Bayesian statistics.* Where you incorporate a certain prior belief in terms of what you think the likelihood of something is before you make a measurement, and that measurement then drives that prior belief into something that's closer to what reality is. *So, I think that Bayesian hypothesis testing is an area that is definitely up and coming, and on the cutting edge of things right now.*" Interview with John Kucera, data scientist on the Adobe Target team. https://blogs.adobe.com/digitalmarketing/personalization/confidence-redefined-overcoming-trust-issues-statistically

Adacus believes that hard-earned consumer attention deserves personalized creative.

By combining 1000+ built-in audience segments, an intuitive decision tree editor and next-gen A/B testing stats that cuts testing time by half, the Adacus Creative-Side Platform enables brands to quickly identify segments and programmatically serve the right message to each segment across their entire media buy.

Learn more at adacus.com