A/B Testing: The Most Powerful Way to Turn Clicks into Customers
By Dan Siroker, Pete Koomen and Cara Harshman

CHAPTER

2

What to Test

Optimization in Five Steps

The hardest part of A/B testing is determining what to test in the first place. Having worked with thousands of customers who do A/B testing every day, one of the most common questions we hear is, "Where do I begin?"

A mistake that some companies make is to start moving a bunch of levers around without clear planning upfront for what they're trying to optimize—and what will be impacted by those changes. It's tempting to just dive in and start changing parts of your homepage, or your product page, or your checkout page, without truly understanding the value that it's generating (or not generating) for your business.

Instead, we advise a purposeful and deliberate five-step process:

Step One: Define success

Step Two: Identify bottlenecks

Step Three: Construct a hypothesis

Step Four: Prioritize

Step Five: Test

This process begins with the most important question of all: *What is the purpose of your site?*

Step One: Define Success

Before you can determine which of your test's variations is the winner, you have to first decide how you're keeping score. To

start A/B testing successfully, you need to answer a specific question: What is your website for? If you could make your website do *one* thing better, what would it do?

If the answer to that question isn't completely clear to you, there's a trick that might help. Imagine the following dialogue:

> ALICE: "What do you want to achieve with A/B testing?"
>
> BOB: "We don't know. We don't know what we want our website to do."
>
> ALICE: "Why don't you take it down?"
>
> BOB: "Of course not! We need our website because it—"

And then Bob has the *aha!* moment that crystallizes his website's *raison d'être*: He can see reasons for the website deeper than "Everyone else has one, so we need one, too."

Defining success in the context of A/B testing involves taking the answer to the question of your site's ultimate purpose and turning it into something more precise: *quantifiable success metrics*. Your success metrics are the specific numbers you hope will be improved by your tests.

It's fairly easy for an e-commerce business to define its success metrics in terms of revenue per visitor (though there are complexities and "gotchas" we'll discuss later), and for a fundraising website to define its success metrics in terms of average donation per visitor. Depending on your business model, defining your success metrics may be trickier.

For instance, Google's search engineers measure what they call *abandonment*, which is when a user leaves a search results page without clicking anything. Abandonment can be bad—perhaps none of the results looked helpful—but it can also be good—perhaps the results page itself was so informative that there was no need to click through to any of the pages.

SITE TYPE	COMMON CONVERSION & AGGREGATE GOALS
E-Commerce A site that sells things for users to purchase online.	• Completed purchase • Each step within the checkout funnel • Products added to cart • Product page views
Media/Content A site focused on article or other content consumption.	• Page views • Articles read • Bounce rate (when measuring within an A/B testing tool, this is often measured by seeing if the user clicked anywhere on the page)
Lead Generation A site that acquires business through name capture.	• Form completion • Clicks to a form page (links may read "Contact Us" for example)
Donation A site aiming to collect donations.	• Form completion • Clicks to a form page (links may read "Send a donation" for example)

FIGURE 2.1 Table of conversion goals by site type.

Figure 2.1 lists some of the most common success metrics for particular site types. Here we've broken websites down into four broad categories.

Part of building out your testing strategy is identifying what constitutes—and does not constitute—a "conversion" for your particular site. In online terms, a conversion is the point at which a visitor takes the desired action on your website. Pinpointing the specific actions you want people to take most on your site and that are most critical to your business will lead you to the tests that have an impact.

Macroconversions, Microconversions, and Vanity Metrics

Author and digital marketing evangelist Avinash Kaushik makes the distinction between what he calls *macroconversions*—the metric

most closely aligned with your site's primary raison d'être, as we discussed earlier—and *microconversions*—the other actions that users take on your site. While microconversions (things like clicking one button in a signup funnel, watching a video, or commenting on a blog post) may not be as immediately valuable or monetizeable as macroconversions, they can provide a tremendous amount of *indirect* value (provided they're not optimized at the expense of macroconversions).

A quick word of caution: sometimes a business can be lured into chasing "vanity metrics" that end up being distractions from the actual goal.

Consider a hypothetical business-to-business (B2B) software company's blog. The marketing team wants the blog to be a hub of thought leadership in their industry. Since they're A/B testing the main site, they decide to start optimizing the blog, too. On the main site, their aim is clear: to use A/B testing to help drive more free trial signups. Defining quantifiable goals for the blog is harder for the team, so they have been unable to define what makes an A/B test *successful*.

For the B2B blog, a vanity metric could be headline clicks. If this is the only piece of data you're using to determine whether the blog is successful, you could be optimizing the wrong thing. Maybe people click headlines because they are shocking, but don't read past them. If all you measure is clicks, you'll never know whether the content of the actual post is good. More telling metrics might be call-to-action clicks, comments, shares, and repeat visits.

Of course, at the end of the day, "thought leadership" is successful only when it results in incremental revenue for the business, but this is very difficult to measure directly. Without clearly defined goals for your site, it's tempting to focus on and optimize for vanity metrics: data that can seem impressive, but doesn't really matter to what you are trying to achieve.

Step Two: Identify Bottlenecks

Once you've determined what your site's quantifiable success metrics are, you can turn your attention to your site analytics and discover where your biggest *bottlenecks* are: the places where your users are dropping off, or the places where you're losing the most momentum in moving users through your desired series of actions.

> DAN: At the Obama campaign in November 2007, before the Iowa caucuses took place, and before our website had much traffic or traction, we did notice one thing by looking at Google Analytics for our user funnel (Figure 2.2).

We had a bunch of people visiting the site—mostly organically—but we also had a very efficient paid marketing campaign. And we were also doing really well getting people to donate once we had their email addresses. The bottleneck was in convincing our site visitors to sign up for our email list (Figure 2.3).

This in turn helped us understand that we had a big opportunity to optimize the email signup step.

The 2008 Obama campaign page generated roughly 10 million email subscriptions, and the lift from the landing page experiment brought in an additional 2.8 million email addresses. Ten percent of the people on our email list volunteered, which meant another 280,000 volunteers. What is perhaps most impressive, and

FIGURE 2.2 Path of the assumed fundraising funnel for the Obama 2008 campaign.

FIGURE 2.3 Path of the actual fundraising funnel at the campaign. High volume of visitors to the site but there's significant dropoff in the email signup step.

most relevant to web businesses, is the lift in terms of the *amount raised*. Because we defined quantifiable success metrics—and knew that we did a great job of raising money from people once we had their email addresses—we had a hunch that if we just got a bunch more email addresses, we'd raise much more money. And sure enough, we raised an additional $57 million.

The Obama example also highlights another equally important part of identifying quantifiable success metrics: agreeing on them. There was a widely held belief inside the campaign at the time that a video would be the most effective media choice for the barackobama.com landing page, and it was only after the team agreed on the definition of *effective* that an objective decision could be made. (We'll explore A/B testing and office decision-making culture further in Chapter 8.)

Step Three: Construct a Hypothesis

Once you've identified what the bottlenecks are in your process, use your understanding of visitor intent to come up with test hypotheses. Consider different forms of qualitative research such as user interviews, feedback forms, or focus groups to gain an understanding of what's going on in users' heads as they interact with your site.

In January 2010, the second-deadliest earthquake ever recorded struck near Léogâne, Haiti. A massive global aid effort began almost instantly, and within days, former presidents Bill Clinton and George W. Bush had established the Clinton Bush Haiti Fund to raise money and support for the relief effort. Time was of the essence and the organization quickly created a simple donation page to collect donations from the millions of visitors the site was seeing every day, thanks to a massive media and press campaign.

The organization was vastly under-resourced at the beginning; they had one extremely overworked IT person who was in charge of the whole operation, and he had barely enough time to make sure the servers were running. (Building a site capable of handling millions of visitors in just a few days is no small feat.) The team at the Clinton Foundation responsible for the website called and said, "Can you help us?"

The donation page was a fertile place for A/B testing to make a big difference: a hastily designed, high-traffic page with a clear conversion goal (Figure 2.4). However, we knew the situation wouldn't last forever, so we really had to hustle in order to make a difference.

We spent three days and nights not only building a series of tests but actually building the scaffolding that would enable us to *run* the tests, effectively constructing the airplane in midflight.

We chose our success metric carefully. We didn't want to optimize for percentage of users making donations, nor for average donation amount, since an increase in one metric might easily be achieved at the expense of the other. The success metric we decided on was dollars per pageview, which was the average amount of money the organization was making per pair of eyeballs seeing this page. We settled on this metric as the best choice to optimize for value to the organization.

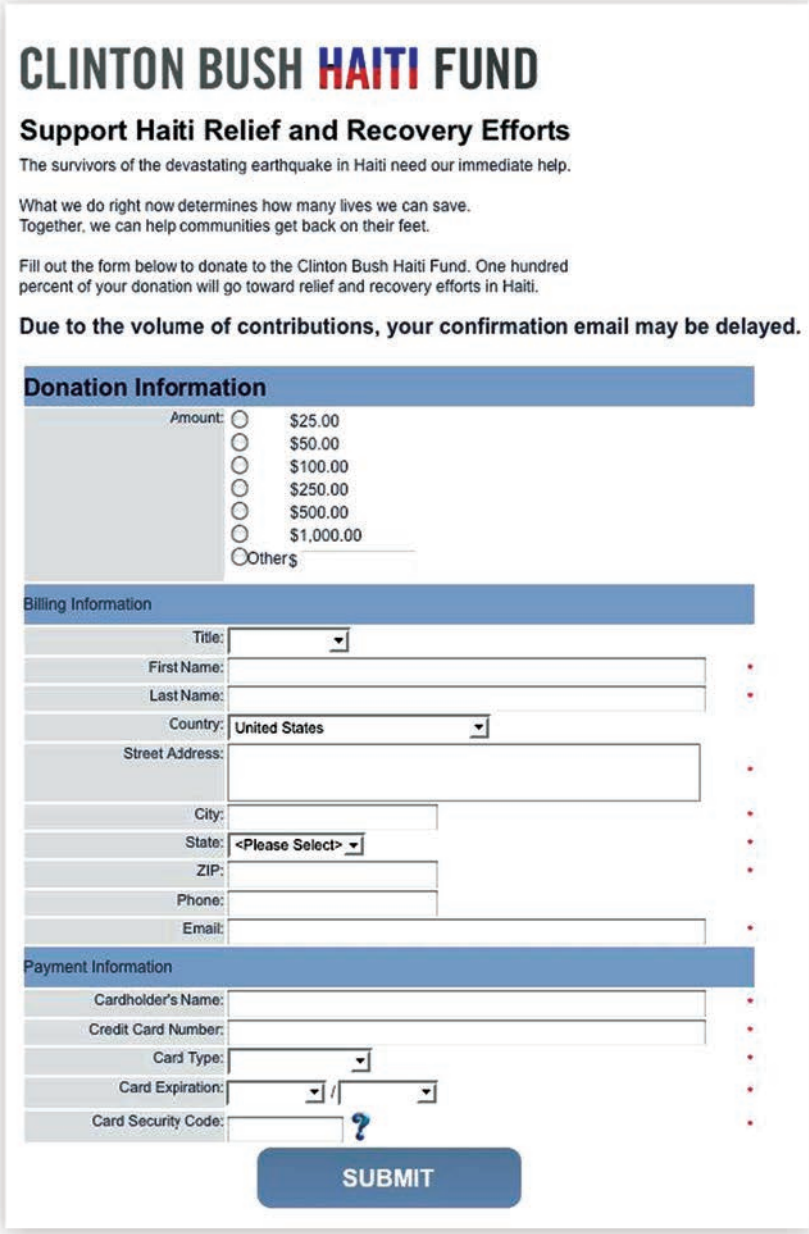

FIGURE 2.4 The initial Clinton Bush Haiti Fund page.

Armed with our metric, we next had to identify where the optimizable areas of the site were. What was the major bottleneck holding back donations? Traffic *to* the site wasn't the problem; we could hardly handle the load. And the site itself was little more than a single page with a donation form, so the bottleneck must be part of the donation page itself.

The initial donation page was essentially a long form, consisting of lots of blank spaces on a white background. We tried to put ourselves inside the visitors' heads, and from their perspective we hypothesized that the form-only page might seem overly abstract. We hypothesized that adding an image of earthquake victims would make the form more concrete and emotive, spurring more visitors to make donations and to make larger donations (Figure 2.5).

Surprisingly, when we tested a variation page with an image against the original, we saw our average donation per pageview go *down*. Here's the point in a testing process where having a hypothesis is critical. Had we simply been trying things at random, we could have easily stopped sending traffic to that variation and never investigated any further. But looking more closely at the page with the image on it led us to a *second* hypothesis: maybe the loss in donations wasn't due to the image itself, but due to the fact that the image was pushing the form down the page ("below the fold"), requiring users to scroll.

What would happen, we hypothesized, if we adopted a two-column layout and put the image *beside* the form? This test would help us make sense of our previous result: whether it was the image itself lowering our metrics, or the layout (Figure 2.6).

It turned out that this layout brought in significantly more donations than not only the failing one-column-with-image layout, but, more importantly and more rewardingly, the original form as well. This new layout (along with several other

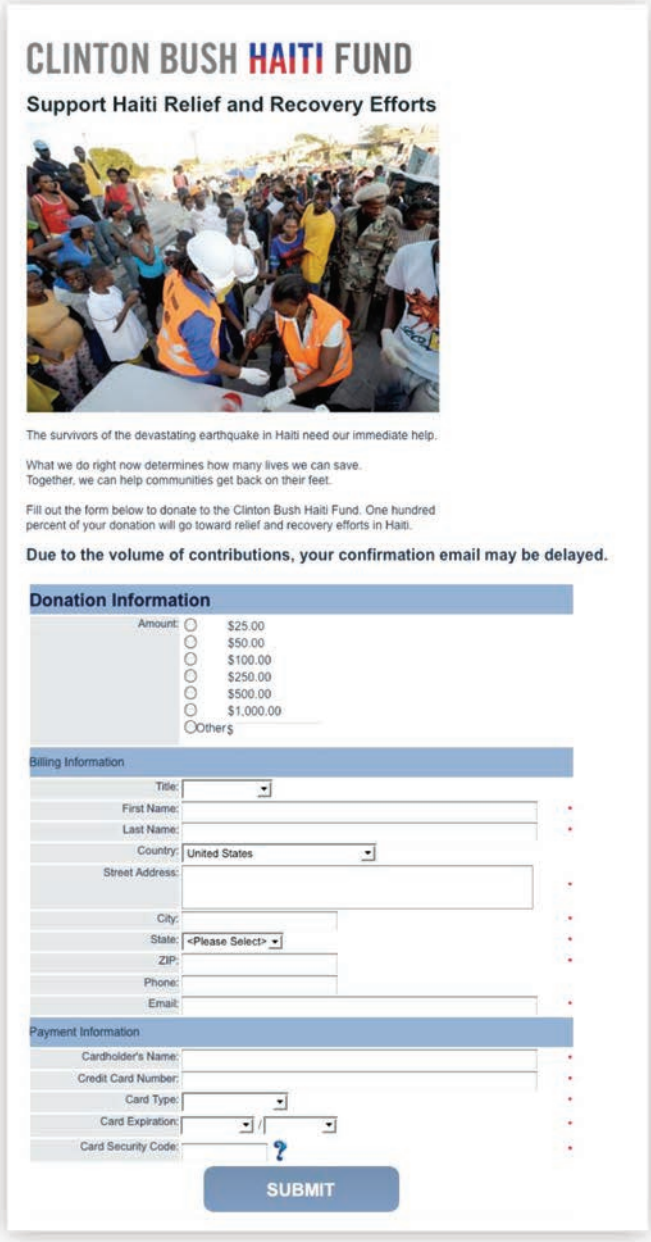

FIGURE 2.5 The Clinton Bush Haiti Fund page with image added.

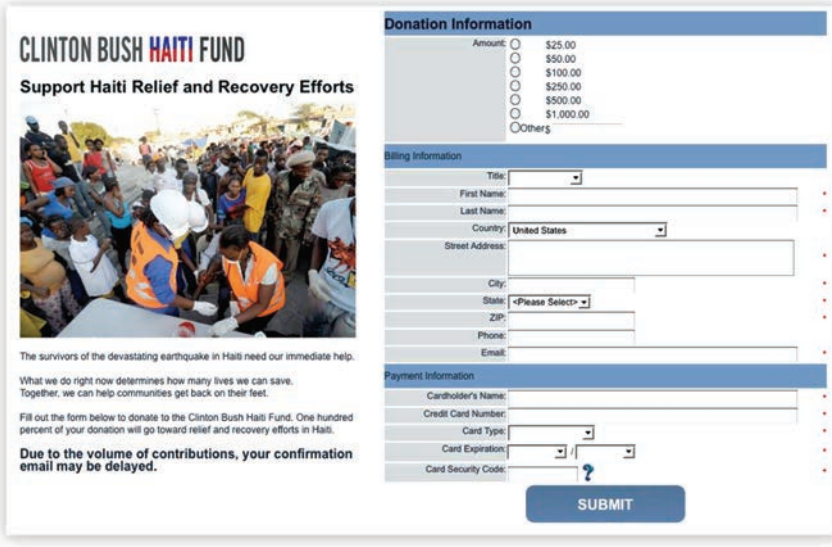

FIGURE 2.6 The Clinton Bush Haiti Fund page with two-column layout.

optimizations that we'll discuss in the chapters to come) led to over a million dollars of additional relief aid for Haiti.

Hypotheses make tests more informative because they provide a specific purpose by helping you hone in on what you are actually trying to determine. If you run an experiment without forming a hypothesis beforehand, you might gather information that's helpful anecdotally while missing the deeper lesson. Experimentation inherently generates more questions than it answers, and when used effectively will always validate or invalidate some hypothesis, thus lending focus to the next round of questions.

"Failed" tests are valuable because they often lead to new hypotheses for why they didn't turn out the way you expected. Generating these hypotheses is sometimes tricky, because visitors behave in complex ways. Regardless of the complexity, however, employing the scientific method in testing will bring you closer to a meaningful understanding of your website's audience.

"With a disaster like the Haiti earthquake, every second counts when it comes to attracting donations, and it goes without saying that every dollar counts," said Marie Ewald, Director of Online Fundraising for the Clinton Foundation. "In less than 48 hours we tested eight versions of the donation page, and through this experiment we were able to generate an additional $1,022,571 in online revenue."

Step Four: Prioritize

Once you've generated hypotheses about user behavior that lead to candidate page variations for testing, you'll need to use your intuition about what's going to have the biggest impact to rank-order your experiments.

"Use ROI [return on investment] to prioritize your tests," says Kyle Rush, who was the Deputy Director of Frontend Web Development at Obama for America. "That's one of the biggest things I've learned in my career."

In a perfect world, you might test absolutely everything, but no team in the real world operates without constraints; your team's attention, budget, time, and also your site's traffic are all finite. These realities make prioritization of testing hypotheses a necessity.

For your very first test, there may be extra considerations, such as wanting to secure the buy-in of others within your organization, or not wanting to try overly elaborate test integrations on a new platform. We'll explore these pragmatic concerns more deeply in Chapter 8. The important thing to bear in mind overall, however, is keeping a sense of your testing priorities. Your projected ROI from each test will itself be derived from a combination of your core success metrics (Step One), the

bottlenecks in your conversion funnel (Step Two), and your hypotheses about your users' behavior (Step Three).

Step Five: Test

All that's left is to run the test. You'll show randomly selected visitors the variation(s) and track how they behave relative to users seeing the current site with respect to the quantifiable success metrics you've determined. (We'll discuss the decision process for choosing the right testing platform in Chapter 7.)

Once the test reaches statistical significance, you'll have your answer. (See Appendix 2 for more information on the mathematics of statistical significance, and the best practices for how long to let a test run.)

Often a completed test yields not only answers, but—as in any other science—more questions. And this cycle of iteration, of exploration and refinement, is exactly where we pick up in Chapter 3.

TL;DR

- You can't pick a winner until you decide how you're keeping score. A/B testing starts with determining **quantifiable success metrics**.
- There are a number of possible **conversion goals**: time on site, pageviews, average order value, revenue per visitor, and so on. Take the time to pick the one that's right for you.
- Site analytics along with your own instincts will suggest **bottlenecks** where you can focus your attention.
- Understanding visitor intent with the help of interviews and usability tests will suggest **hypotheses** about what to change and how.
- **Prioritize** your experiments based on your prediction of their impact.
- **Start testing**, and continue until you begin to see diminishing returns.